# Cross-language relevance assessment and task context

Jussi Karlgren and Preben Hansen
Swedish Institute of Computer Science, SICS
Box 1263, SE-164 29 Kista, Sweden
{jussi, preben}@sics.se

## Abstract

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.

## 1. Cross-linguality and reading

### 1.1 People are naturally multi-lingual

For people in cultures all around the world competence in more than one language is quite common and the European cultural area is typical in that respect. Many people, especially those engaged in intellectual activities are familiar with more than one language and have some acquaintance with several.

### 1.2 People are good at making relevance assessments

Information access systems deliver results which on a good day hold up to forty per cent relevant items. It is up to the reader to winnow out the good stuff from the bad.

We know that readers are excellent at making relevance assessments for texts. Both assessment efficiency and precision are very impressive. But how we go about it we know very little about. Practice seems to improve both assessment speed, assessment precision, and assessor confidence, but what features a reader focuses on and how they are combined has not been studied in any great detail.

### 1.3 Linguistic competence is a continuum

Languages are tools tied to tasks. For any one task, typically people have one language they prefer to perform it in. In general, while people may have working knowledge of more than one language, it is not common for people to have equal competence in many; the first language, or the school language, or the workplace language will tend to be stronger for whatever task they are engaged in. Linguistic competence is not a binary matter: people know a language to some extent, greater or lesser. What bits of competence are important in any given situation is an ongoing discussion in the field of language teaching – we will here concentrate on some aspects of reading, related to situation, task, and domain.

### 1.4 Assessing relevance in a strange language is hard – and important

We do know that reading about strange things in strange genres takes more time than familiar genres, and that reading a language we do not know well is hard work, and something we only attempt if we believe it is worth the effort.

Judging trustworthiness and usefulness of documents in a foreign language is difficult and a noticeably less reliable process than doing it in a language and cultural context we are familiar with.

These starting points have immediate ramifications for the design of cross-lingual and multi-lingual information access systems. Presenting large numbers of documents to users if it is likely they will not be able to determine their usefulness is a waste at best and a trustworthiness and reliability risk at worst.

### 1.5 Finding out more – does language make a difference?

We need more data about reading and related processes. To find out more we set up an experiment where Swedish-speaking subjects, fluent in English as determined by self-report, were presented with retrieval results both languages, and given the task of rating the results by relevance. Our hypotheses were that results for a foreign language would be more time-consuming and less competent than those for the first language.

### 1.6 Task-based approach to query construction and relevance assessment

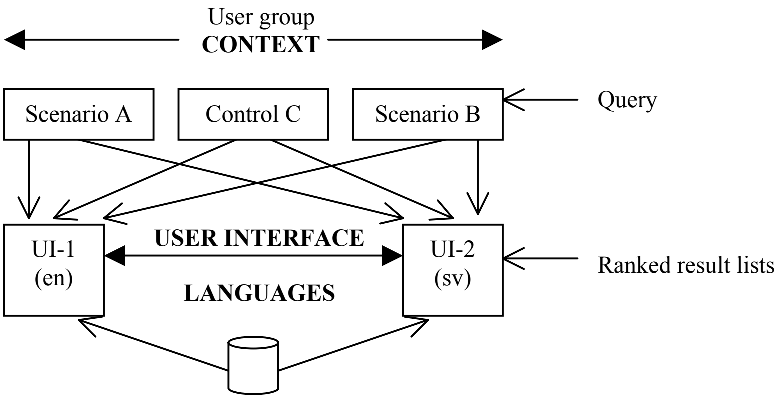
Generally, topicality has been the main criteria for relevance in information retrieval experiments. Our approach suggests that other criteria may come into play, especially criteria related to the task and domain at hand. For interactive information retrieval experiments, we propose to expand the original query with information about context. 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.

## 2. Experiment

### 2.1 Set-up

- *Participants*: 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.
- *Scenario*: Each scenario had 4 participants.
- *Language*. 2 languages were used: English and Swedish.
- *Queries*. The four CLEF queries used in this year's interactive track were used in both languages: queries 53, 56, 65, and 80. Query 86 was used for a practice run.
- *Result list.* Sets of ranked result lists of length between one and two hundred 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.
- *Presentation.* The ranked lists were presented to the participants, varied by order and language (cf. Table 1) in a simulated search interface.
- *System.* The experiment infrastructure was built using HTTP and was deployed over the WWW. The canned ranked results were put up as html pages and linked to the actual documents, which were displayed with four buttons to be used for the relevance ranking. A simple cgi-bin based logging tool noted the relevance assessment made and the time taken to make the assessment after display of the document.
- *Questionnaires*. The participants filled out questionnaires at various points in the study. The data was collected either by semi-structured questions or measured by a Likert scale of 1 to 5 or 1 to 3.
- *Relevance categories.* The participants could in the interface indicate for each document one of four assessments: "not relevant" "somewhat relevant", "relevant", and "don't know".

Figure 1. Task and scenario-based experiment design



### 2.2 Simulated Domain and Work-Task Scenarios

In this study we use the Simulated Domain and Work-Task Scenario (SDWS) methodology, an evaluation methodology with simulated contexts that include description of domains and work-tasks. The method is an extension of the notion of simulated work-tasks (Brajnic et al., 1995; Borlund, 2000; Ruthven et. al., 2002) among others. Borlund and Ruthven enhanced the context of standard queries using two fields with descriptive information. We extend this design to include a domain description and a general work-task description. The goal of the method is to give the experimental query a context closer to a real-life information-seeking situation. In this way, the SDWS would allow the user a) a broader understanding of the situation, and b) a subjective interpretation of the relevance.

Constructing a SDWS query within a context was done by creating two levels of description (cf. Figure 2): a general description including a short description of the *domain* and a short description of general *work-tasks* or routines that are performed. The next level contains a situational description including the *topic* of the query (in this case the I-clef query) and a *search task* description, which also include parts of the description field of the actual I-clef query (cf. appendix A for a SDWS for query CO53).

Figure 2: Design of the simulated Domain and Work-Task scenario

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

### 2.3 Procedure

The participants were asked to answer some initial questions. After that, participants in groups A or B were asked to read through a workplace scenario carefully and try to act within the assigned scenario as well as possible. Then participants were asked to read through the first work-task related query and to assess the ranked list for it pursuant time constraints as per the scenario, or in the case of group C, to keep the time about constant around fifteen to twenty minutes per query. After the assessment participants were asked to answer a fixed set of questions related to the query and the work task. This fixed set of questions was repeated after each of the four queries. Finally, after the last query, participants were asked to answer a last set of questions.

Table 1: Matrix of scenarios, queries and languages used in experiment

| Scenario A | | | Scenario B | | | Scenario C (control group) | | |
|---|---|---|---|---|---|---|---|---|
| User1 | L:SE Q:1+3 | L:EN Q:2+4 | User5 | L:SE Q:1+3 | L:EN Q:2+4 | User9 | L:SE Q:1+3 | L:EN Q:2+4 |
| User2 | L:EN Q:1+3 | L:SE Q:2+4 | User6 | L:EN Q:1+3 | L:SE Q:2+4 | User10 | L:EN Q:1+3 | L:SE Q:2+4 |
| User3 | L:SE Q:3+1 | L:EN Q:4+2 | User7 | L:SE Q:3+1 | L:EN Q:4+2 | User11 | L:SE Q:3+1 | L:EN Q:4+2 |
| User4 | L:EN Q:3+1 | L:SE Q:4+2 | User8 | L:EN Q:3+1 | L:SE Q:4+2 | User12 | L:EN Q:3+1 | L:SE Q:4+2 |

### 2.4 Participant

The 12 participants in this study had a variety of academic and professional backgrounds. 5 participants were male and 7 female, with an average age of 36,5. The participants had an overall high experience searching web-based search engines such as Google (4,33) and an overall low experience in searching commercial databases (2,16) and using machine translation tools such as Babel-fish (2.00). 2/3 of the participants used some kind of search engine 1-2 times every day. Average on overall knowledge in English was 4,25 (see app. B for a full version and table of the pre-questionnaire). Note that this information is based on the participants' own subjective judgments.

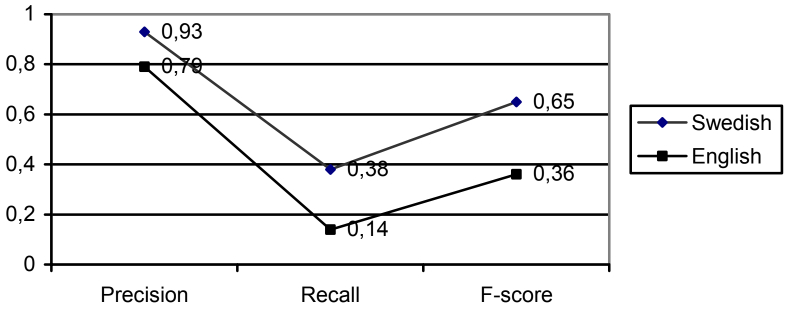
## 3. Results

### 3.1 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 (cf. Figure 2; all differences between English and Swedish 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. In general, the precision is reasonably high for both languages, which can be taken to indicate that participants went through the list and found most relevant documents in the presented list.

All documents are very short. The Swedish documents are from a wire service and the English documents from a newspaper. The average length of an English article is over seven hundred words, whereas the Swedish articles are of an average length of just over four hundred. The difference in averages is partially due to the English average being highly skewed from a few very long feature articles, a genre almost entirely missing from the Swedish corpus. 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

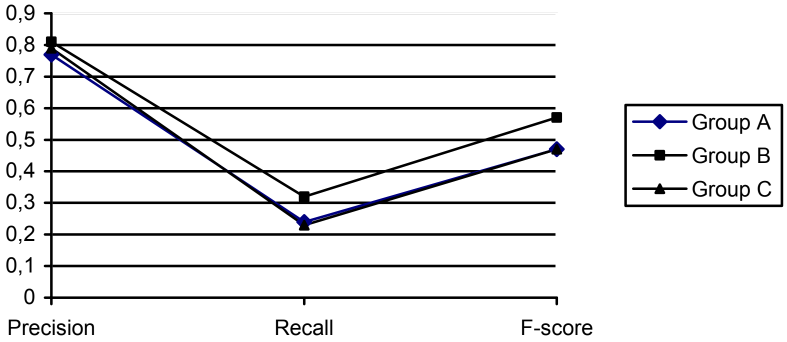
Figure 2: Retrieval results



### 3.2 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 or 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 result by task



### 3.3 Relevance judgment aspects

We assumed that aspects of the relevance judgment taken into account would extend beyond traditional topicality. In order to see if aspects other than topicality were taken into account, 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. We present the results for groups A and B in Table 2. Merged, the two groups used the domain related aspect in 12% of the cases, the task related aspect in 46% of the cases, and the topic-related aspect in 42% of the cases. All observations were done over all four i-CLEF queries given to the participants. 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.

Table 2: Type of relevance judgement aspect by scenario, for both languages combined.

| | Swedish and English | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Group A | | | | Group B | | | | | | | |
| | R1 | R2 | R3 | SUM | R1 | R2 | R3 | SUM | R1 | R2 | R3 | *TSUM* |
| Q53 | 1 | 2 | 1 | *4* | | 2 | 3 | 5 | 1 | 4 | 4 | *9* |
| Q56 | 2 | 2 | 4 | *8* | | 4 | 1 | 5 | 2 | 6 | 5 | *13* |
| Q65 | 1 | 3 | 3 | *7* | | 2 | 1 | 3 | 1 | 5 | 4 | *10* |
| Q80 | 1 | 2 | 3 | *6* | | 3 | 2 | 5 | 1 | 5 | 5 | *11* |
| TSUM | *5* | *9* | *11* | *25* | *0* | *11* | *7* | *18* | *5* | *20* | *18* | *43* |
| Mean | | | | *1,56* | | | | *1,12* | | | | *1,34* |

Legend: n=4 n=4 n=8

R1= Relevance judgement aspects related to the task domain (translator and news agents)

R2= Relevance judgement aspects related to the task given to the participant

R3 = Relevance judgement aspects related to the topic of the query

## 3. Discussion

The results are quite convincing. 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. 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. 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. This is but a first step in this direction; we intend to pursue this avenue of inquiry further, and investigate its effects on design. Specifically, during the coming year we will investigate if adding more information to the interface will improve results for the foreign language assessment task.

## Acknowledgments

We thank Tidningarnas Telegrambyrå AB, Stockholm, for providing us with the Swedish text collection, and Heikki Keskustalo, University of Tampere, Johan Carlberger and Hercules Dalianis, Euroseek AB, for kind support in producing the ranked lists of documents. Furthermore, we gratefully acknowledge funding provided to us the European Commission under contracts IST-2000-29452 (DUMAS) and IST-2000-25310 (CLARITY). And finally we thank our patient subjects for the time they spent reading really old news.

## References

Borlund, P. (2000*). Evaluation of Interactive Information Retrieval Systems. Doctoral dissertation*. Åbo, Finland: Åbo Academi

Hansen, P., & Järvelin, K. (2000). The Information Seeking and Retrieval process at the Swedish Patent- and Registration Office. Moving from Lab-based to real life work-task environment. Proceedings of the ACM-SIGIR 2000 Workshop on Patent Retrieval, Athens, Greece, July 28, 2000, pp. 43-53.

Ruthven, I., Lalmas, M. and van Rijsbergen, K. (2002). Ranking Expansion Terms with Partial and Ostensive Evidence. Proceedings of the Fourth International Conference on Conceptions of Library and Information Science – CoLIS4, Seattle, USA, July, 2002, pp. 199-220.

# Appendix A

### The SDWS framework description

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

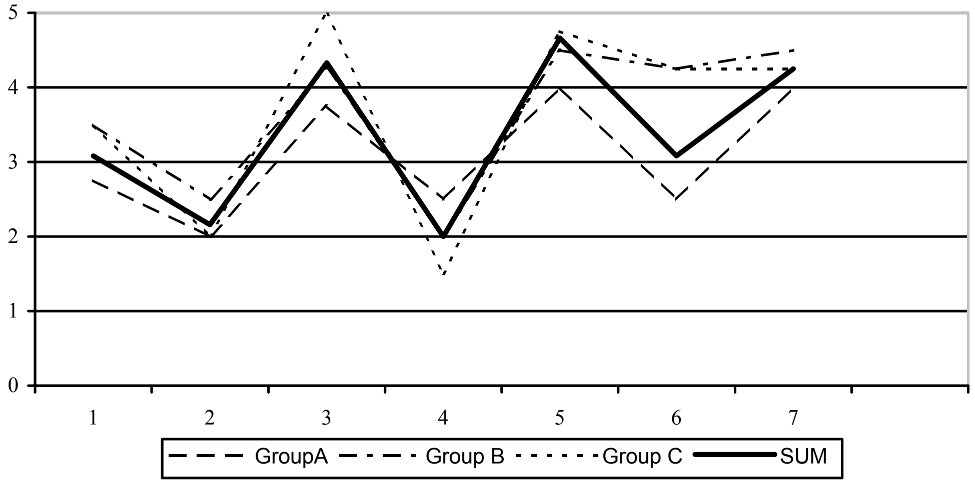
# Appendix B

### The pre-questionnaire

Table 3: Background competence

| What is your experience in… | No experience<br>1 | 2 | Some experience<br>3 | 4 | Extensive experience<br>5 |
|---|---|---|---|---|---|
| Searching online library catalogues? | | | | | |
| Searching commercial databases (such as Dialog) | | | | | |
| Searching Internet-based search engines such as Google | | | | | |
| Using tools for machine translation (such as Babelfish) | | | | | |
| | Never<br>1 | 1-2 times a year<br>2 | 1-2 times a month<br>3 | 1-2 times a wk<br>4 | 1-2 times a day<br>5 |
| How often do you use any kind of search engine? | | | | | |
| | Strongly disagree<br>1 | Disagree<br>2 | Neutral<br>3 | Agree<br>4 | Strongly agree<br>5 |
| I like searching for information | | | | | |
| | None | Poor | Fair | Good | Very good |
| My reading skills in English ... | | | | | |

Figure 6: Background competence


GroupA
Group B
Group C
SUM