Measuring a cross language image retrieval system

Mark Sanderson, Paul Clough, Catherine Paterson, Wai Tung Lo
Department of Information Studies, University of Sheffield, S1 4DP, UK

Abstract

Cross language information retrieval is a field of study that has received significant research attention, resulting in systems that despite the errors of automatic translation (from query to document), on average, produce relatively good retrieval results. Traditionally, most work has focussed on retrieval from sets of newspaper articles; however, other forms of collection are being searched: one example being the cross language retrieval of images by text caption. Limited past work has established, through test collection evaluation, that as with traditional CLIR, image CLIR is effective. This paper presents two studies that start to establish the usability of such a system: first, a test collection-based examination, which avoids traditional measures of effectiveness, is described and results from it are discussed; second, a preliminary usability study of a working cross language image retrieval system is presented. Together the examinations show that, in general, searching for images captioned in a language unknown to a searcher is usable.

Introduction

A great deal of research is currently conducted in the field of Cross Language Information Retrieval, where documents written in one language (referred to as the *target* language) are retrieved by a query written in another (the *source* language). Until now, most work has concentrated on locating or creating the resources and methods that automatically transform a user's query into the language of the documents. With the right approach, CLIR systems are relatively accurate: managing to achieve retrieval effectiveness that is only marginally degraded from the effectiveness achieved had the query been manually translated (referred to as monolingual retrieval). For example, Ballesteros (1998) achieved CLIR effectiveness at 90% of monolingual retrieval.

One area of CLIR research that has received almost no attention is retrieving from collections where text is used only to describe the collection objects, and the object's relevance to a query are hopefully clear to anyone regardless of their foreign language skills. One such collection is a picture archive where each image is described by a text caption. Retrieval from such an archive presents a number of challenges and opportunities. The challenges come from matching queries to the typically short descriptions associated with each image. The opportunities derive from the unusual situation of potentially building a CLIR system that a large number people may wish to use. For any vendor of an image library, use of CLIR offers the opportunity of expanding the range of potential searchers of (even purchasers from) their library.

In the rest of this document, previous work in cross language and in image retrieval is described, along with past image CLIR research. This is followed by the design and results of the first test collection based study, which led onto a preliminary targeted usability experiment, which is also outlined. Finally, overall conclusions and avenues for future work are provided.

Previous work

Because there is little past work investigating cross language image retrieval, this Section will first review the two component research areas separately: image retrieval by associated text and cross language IR; followed by an examination of the proposals and occasional attempts at performing image CLIR.

Image retrieval

Retrieval of images by text queries matched against associated text has long been researched. As part of his PhD investigating multimedia retrieval, Dunlop examined such an approach to image retrieval (1993). The ideas in this work were later extended to a study of image retrieval from art gallery Web sites by Harmandas et al (1997), who showed that associated text was well suited for retrieval over a range of query types. At the same conference, research was presented on the successful use of a thesaurus to expand the text of image captions (Aslandogan et al 1997). More recently, research in combining content-based image retrieval with caption text has been explored in the work of Chen et al (1999).

Cross language retrieval

Most CLIR research effort has focussed on locating and exploiting translation resources. Successful methods centre on use of bilingual dictionaries, machine translation systems, or so-called parallel corpora, where a large set of documents written in one language are translated into another and word translations are derived from their texts. With reasonably accurate translation, effective cross language retrieval is possible. This is being confirmed in the large-scale evaluation exercises within TREC[1] and CLEF[2]. A thorough review of this aspect of CLIR research can be found in Gollins (2000).

Cross language image retrieval

The potential utility of image CLIR has been known about for sometime: both Oard (1997) and Jones (2001) discussed the topic. However, only a few examples of image CLIR research exist.

1. The IR Game system built at Tampere University (Sormunen, 1998) offered Finish/English cross language retrieval from an image archive, images were ranked using a best match search. However, little has been written about this system.
2. The European Visual Archive (EVA[3]) offered English/Dutch/German cross language searching of 17,000 historical photographs indexed by a standard set of 6,000 controlled terms. Searching is restricted to Boolean searching.
3. Flank (2002) reported a small scale test collection style study of image CLIR using a collection of 400,000 images captioned in English and ten queries translated into three languages. She concluded from her results that image CLIR was effective. However, the conclusions were based on a collection that by most standards is too small: Voorhees (1998) emphasised the importance of working with test collections that have at least 25 queries; any fewer and results derived maybe unreliable.
4. In 2002, the imageCLEF track of the CLEF (Cross Language Evaluation Forum) exercise was established with the release of one of the first publicly available image test collections: consisting of approximately 30,000 images (from a digital collection held at St. Andrews University in Scotland) and fifty queries. The collection was used by four research groups. Working across a number of European languages – and in the case of one research group, Chinese – it was shown that cross language retrieval was operating at somewhere between 50% and 78% of monolingual retrieval (Clough, 2003), confirming Flank's conclusion that image CLIR can be made to work.

From these past works, one might conclude that image CLIR is feasible; however, the conclusion would be based on test collection evaluation alone. It is also necessary to consider if user will be able to judge retrieved images as relevant.

Judging images for relevance

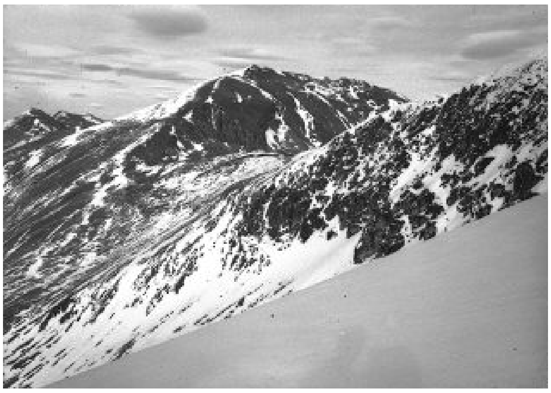
In past works where image CLIR has been suggested as an application of cross language technologies (Oard, 1997 and Jones, 2001), there has been an assumption that users are generally capable of judging the relevance of images simply by observing them: reading caption text is not necessary. It is to be anticipated, however that there will be limits to such a capability. For example, in the two images shown below, most would agree the first was relevant to a query for mountain scenery without having to read any caption; however, for a query requiring a picture of London Bridge, the only way most searchers would know if the second image was relevant or not was by reading an associated caption. This problem was highlighted in an extensive study of image retrieval by Choi and Rasmussen (2002) who showed user's judgement of image relevance was often altered once they read the image's caption (p. 704). Aware of this issue, Flank (2002) in her small test collection-based experiment, limited the

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

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

[3] http://www.eva-eu.org/

scope of the ten queries used to a generic set of topics that she believed most users would be able to judge image relevance from.



It appears that in order for an image CLIR system to work well, it will be necessary to include a translation of captions in order to provide users as much information as possible to judge relevance. In text document CLIR, Resnik reported his early work on so-called gists of documents (1997), however evaluation of the system was such that it is hard to speculate on users' abilities to judge relevance from the gists. This ability was more thoroughly assessed by Gonzlao et al (2004) who reported the results of the CLEF interactive track. Here it was shown that users were capable of determining the relevance of retrieved documents accurately after they were automatically translated in some manner.

Problems with test collection measures

Effectiveness of retrieval system is almost always measured on a test collection using either average precision measured across a series of recall values or at a fixed rank. It would appear that this is done as these measures are commonly assumed to be a reasonable model of a user's view of the effectiveness of a retrieval system: the higher the average precision, the more satisfied users will be. How true this view is appears, perhaps surprisingly, not to have been tested as thoroughly as one might expect. For example, techniques exist, such as pseudo relevance feedback, that consistently produce higher average precision than baseline systems, implying that they would be preferred by users, but they are rarely deployed in actual IR systems. Such a contradiction suggests that, in this case at least, average precision is not reflecting user preferences. In CLIR, effectiveness is measured as the percentage reduction in average precision from that achieved with monolingual queries. With the result from Clough & Sanderson (a lower bound of 50% of monolingual), one may conclude that for every ten relevant documents resulting from a monolingual search, five are retrieved for every query tested. However, the consistency of the reduction in effectiveness is uneven, as a study of individual queries reveals.

In order to better understand the unevenness for image CLIR, a test collection evaluation was conducted where the measures of average precision and precision at rank ten were dropped in favour of a measure perceived to be more "user centred". Before describing the experiment, the collection that searching was conducted on is first outlined.

The collection

Selecting a suitable collection for image retrieval was a non-trivial task. Not only was a "large" collection of images required, but also a collection with captions of high quality to facilitate text-based retrieval methods. Additional issues involving copyright were also encountered as typically photographs and images have a potentially high marketable value, thereby restricting permissible distribution. A library that was willing to release its collection was eventually found. St. Andrews University[4] holds one of the largest and most important collections of historic photographs in Scotland, exceeding over 300,000 photographs from a number of well-known Scottish photographers (Reid, 1999). A cross-section of approximately 30,000 images from the main collection was digitised to enable public access to the collection via a web interface. Permission was sought and granted by St. Andrews Library to downloaded and distribute the collection for research use.

[4] http://www-library.st-andrews.ac.uk/

This collection was used as the basis for ImageCLEF and the experiments described here. The collection consists of 28,133 images (a 368 by 234 pixel large version along with a 120 by 76 thumbnail[5]) and captions. The majority (82%) of images are in black and white ranging between the years 1832 and 1992 (with a mean year of 1920). Images and captions of varying styles, presentation, and quality exist in the collection. The figure below shows an example image and caption from the collection. The captions consist of data in a semi-structured format added manually by domain experts at St. Andrews University. The caption contains eight fields, the most important being the description, which is a grammatical sentence of around fifteen words. The captions exist only in British English, and the language tends to contain colloquial expressions.

Title	Old Tom Morris, golfer, St Andrews
Short title	Old Tom Morris, golfer
Location	Fife, Scotland
Description	Portrait of bearded elderly man in tweed jacket, waistcoat with watch chain and flat cap, hand in pockets; painted backdrop.
Date	ca.1900
Photographer	John Fairweather
Categories	[golf - general], [identified male], [St. Andrews Portraits], [Collection - G M Cowie]
Notes	GMC-F202 pc/BIOG: Tom Morris (1821-1908) Golf ball and club maker before turning professional, later Custodian of the Links, St Andrews; golfer and four times winner of the Open Championship; father of Young Tom Morris (1851-1875). DETAIL: Studio portrait.

Test collection study

As part of preparations for the formation of the imageCLEF collection (Clough and Sanderson, 2003), a preliminary evaluation of image CLIR was conducted on the St. Andrews collection of images, the full details of which are described by Paterson (2002). A set of fifty queries was created in English, which were manually translated into Portuguese and German. The queries were then translated back into English using AltaVista's Babel Fish[6] and submitted to an in-house searching system which uses a BM25 weighting scheme, searching the full text captions of the collection. With similar results to the imageCLEF experiments described above, the conclusions of the study revealed that CLIR of German and Portuguese retrieved respectively between 55% and 67% of the relevant documents retrieved by monolingual searching. As with imageCLEF, the results here were measured with a form of average precision, in this case, precision at rank ten. Average precision measures, across queries, the density of relevant documents measured at certain fixed positions: either at a fixed rank position, as used here; or measured at standard levels of recall as is commonly found in recall/precision graphs. As already described in this paper, it is not clear how well average precision models user preferences. Consequently, an alternative measure was sought. In an attempt to derive one, a consideration of what was important to searchers when using retrieval systems was conducted. Three cases were considered and user reactions conceived:

1. **Returning one relevant document**: at a minimum, users will be content with finding a single relevant item returned by their query. If the cross language system offers relevance feedback, users are likely to be able to locate more relevant items through use of that process;
2. **Returning more than one relevant document**: although additional user satisfaction would inevitably result from a larger number of relevant items being retrieved, the benefits to users of this aspect is not as important as the detriments of the following situation;
3. **Returning no relevant documents**: a great deal of user dissatisfaction would be associated with no relevant documents being retrieved. This may be particularly important for cross language retrieval, as reformulation of a query is likely to be harder than with monolingual searching.

The conclusion of this consideration was that emphasising user dissatisfaction at failed queries in the effectiveness measure is the priority. Consequently, it was decided that simply counting the number of

[5] Much higher resolution images are retained by St. Andrews and are not part of the collection.
[6] http://babelfish.altavista.com/, service used in the summer of 2002.

queries resulting in at least one relevant image being returned in the top ten would be an effective measure of user satisfaction as it would better reflect user preferences over the range of searching situations as well as being a clear and simple measure to understand. The following table shows the effectiveness of the retrieval system across the three languages as measured by successful queries and, for comparison, by average precision at rank ten.

1	2	3	4	5	6
Query language	**Query retrieves ≥1 relevant**	**Absolute %**	**Relative %**	**Precision at rank 10, %**	**Relative %**
English	48	96	100	49	100
German	35	70	73	27	55
Portuguese	39	78	81	33	67

It is our contention that in comparison to averaged precision (column 5), the measure of successful queries (column 2) provides a clearer notion of the user experience when searching: to see that two in fifty queries fail with English in comparison to the fifteen in fifty for German is more tangible (we would claim) than observing the density of relevant documents in the top 10 reduced from 49% to 27%. It is also worth noting that the reduction in effectiveness, measured by successful queries (column 4) showed a smaller reduction in effectiveness from monolingual than that measured with precision (column 6). From the data in the table, we draw two conclusions

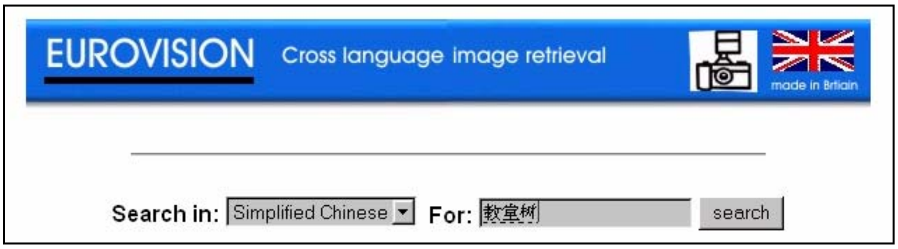
1. When considering the number of failed queries in cross language retrieval, it can be seen that a significant number of queries fail.
2. Although the failure rate is high, it is notable that although the density of relevant images retrieved drops considerably, from monolingual to cross language the number of "failed queries" does not drop as much.

The overall conclusion of this experiment is that as with previous research, image cross language retrieval is workable, however, by analysing the number of queries that return no relevant images in the top 10, it is clear that users of such a system will be faced with the problem of having to deal with failed queries on a reasonably frequent basis.

To further understanding of image CLIR, however, it was decided a usability test be conducted to examine how users may cope with querying for images across languages. Before describing the test, however, the image CLIR system that the test was conducted on is first described.

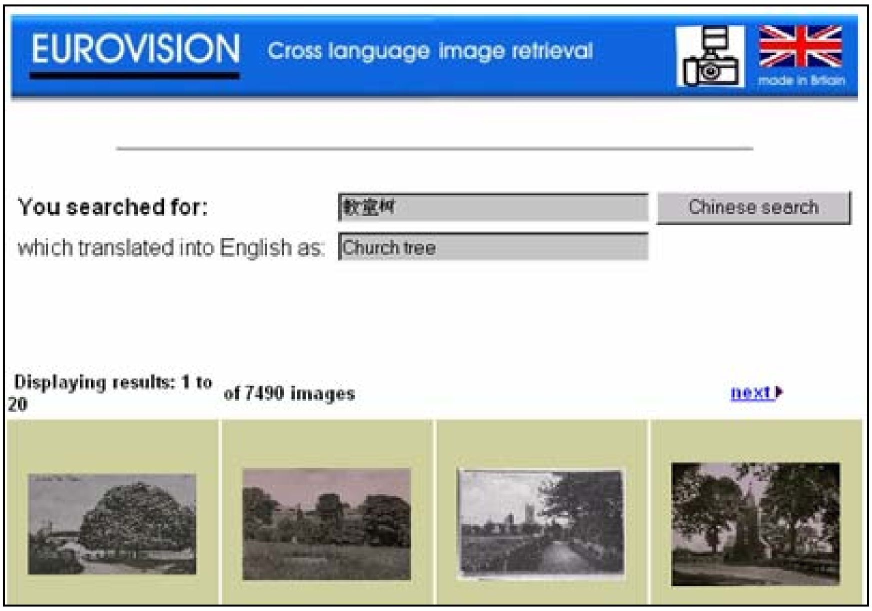
The image CLIR system

The system tested was an image caption-based text retrieval system created "in-house" by the authors. It offers a straightforward search engine-like interface, where users enter their search terms in their native language (in the figure below, Chinese).



Search in: Simplified Chinese For: 教堂树

As with the earlier test collection study described above, the AltaVista Babel Fish translation system was accessed when translating a user's query from the source to the target language (i.e. Chinese to English). Retrieval on the source language version of the query was performed by an IR system searching on all parts of the image captions, using BM25 to rank images. As can be seen below, both the original query and its translation (correct in this case) are displayed along with the retrieved images. The retrieved set is shown as a table of thumbnails grouped in batches of twenty. To avoid the risk of over using the Babel Fish service, translation of image captions was not performed.



You searched for: 教堂树

which translated into English as: Church tree

Displaying results: 1 to 20 of 7490 images

Users were free to re-enter or modify queries as they preferred.

The experiment

The usability experiment was set up as a preliminary exploration of how capable users were at cross language image searching. It was decided to set a series of *known item* searching tasks. Subjects were shown an image (without any caption text) and asked to find that image within the collection. This caused them to try to find a form of query words that would help them locate the image. The fact that captions of retrieved images in the users' language were not available was less of a problem as judgement of relevance of the known item could be made without recourse to caption text. It was expected that searchers would be willing to put more effort into searching and reformulation than a more classic IR "find images relevant to topic x" form of task as they knew the image was within the collection and knew they hadn't succeeded until they found it.

The subjects

As the collection being searched was captioned in English, the queries were written in (and the searchers had to be fluent in) a different language. With such a collection, it would be preferable for the searchers not to know English. Locating such people within the UK proved to be too hard a task. However, it was possible to locate a large number of bilingual people. For the usability experiment, eighteen native Chinese language speakers were recruited. They were students at the University of Sheffield taking a postgraduate Masters course; each was paid £15 for participating. As with any usability test, the time one can take with each subject was limited by the amount of time someone can reasonably be expected to continually search before tiring. Therefore, each searcher spent approximately one hour conducting the experiment, completing pre and post test questionnaires and being interviewed.

The pre-test questionnaire established that the subjects felt they were good at searching; most of them making at least weekly access to either a search engine or electronic library catalogue. A problem, for the experimental setup with these subjects was that by their own admission their command of English was good (66%) or fair (22%); only two (12%) regarding their English as basic. With such language skills, if the subjects viewed the English captions of the retrieved images, they would become frustrated by having to search in Chinese through a translation system, perhaps preferring to re-formulate their queries in English. Therefore, it was decided that the captions of the retrieved images would be removed from display of the interface. As the task chosen for users was a known item task, captions were not needed to help with judging relevance, therefore, their removal for this type of experiment was judged not to be problematic.

Experiment

A series of images were selected to give users a range of difficulty of locating a known item. The images were of a bridge, a ship, a Tapir, an empty street, and a person:

- A bridge, and a ship – these images were expected to be relatively easy to query for;
- the Tapir, less so, as users may not know the name of this animal;
- the person and street scene would be hard to locate without knowing the name of the street or the town it was in.

No time limit was placed on the users' execution of each known item search. The number of queries issued, the number of search result pages viewed (in total), the time taken, and the success rate of the experimental subjects was logged.

Results

As expected, across the five known items, success varied: for three of the five images nearly all users were able to locate the known item. The image of the bridge was hard to locate as the collection holds a pictures of a great many. Users consequently searched through a large number of result pages (eighteen on average). Determining an appropriate query to retrieve the street scene image proved almost impossible for users, although a number of queries (on average six) were tried by the experimental subjects in their attempts to locate it; only one user succeeded. Most users located the image of the person, though as with the bridge, many images of people are to be found in the St. Andrews collection.

Image	*Bridge*	*Ship*	*Tapir*	*Street*	*Person*	**Average**
Queries	3.4	2.8	1.8	6	4.3	3.66
Pages viewed	18	2.6	1	22	18.7	12.46
Time (min.)	6.2	1.7	1.4	7.0	7.8	4.82
Success (%)	88.9	100	100	5.6	66.7	72.24

A number of conclusions about image CLIR were drawn from the results of this preliminary test:

- Users appear to be relatively successful in known item searching within an image CLIR system. The tasks set were not trivial (especially in the case of the final two images) and one would not expect users to be 100% successful even if they were searching in English. It is to be hoped that the success observed here will transfer to more general image CLIR tasks.
- Users appear to be willing to re-submit and re-formulate queries to a CLIR system in order to locate an item.
- Users are willing to look through a large number of retrieved images in order to find what they are seeking.

The overall conclusion of the work was that users are capable of searching for images in a CLIR system with a relatively high degree of success.

Conclusions and future work

Two experiments were described where the feasibility of image CLIR was examined. First, a test collection-based study explored a different means of measuring the effectiveness of a CLIR system. It was argued that the measure better illustrated the problems with CLIR, namely queries that fail to retrieve any relevant images. Second, a small usability study of a working image searching system was tested with users querying in a language different from that of the image captions. Here, it was concluded that users were more than capable of searching for items in a collection; a conclusion that bodes well for CLIR when applied to image collections.

For future work, the effectiveness of automatic translation of image captions will be examined and consequently a wider ranging usability test will be conducted to broaden the set of tasks users are observed completing. A re-examination of existing CLIR research is planned where measuring and comparing past results using the failed queries statistic will be conducted.

Acknowledgements

This research was supported in part by the EPSRC fast track grant Eurovision (GR/R56778/01).

References

Aslandogan, Y.A., Thier, C., Yu, C.T., Zou, J., Rishe, N. (1997) Using Semantic Contents and WordNet in Image Retrieval in *Proceedings of ACM SIGIR '97*, 286-295.
Ballesteros, L., Croft, W.B. (1998): Resolving ambiguity for cross-language retrieval, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, J. Zobel (eds.): 64-71

Chen, F., Gargi, U., Niles, L., Schuetze, H. Multi-Modal Browsing of Images in Web Documents, *Proceedings of SPIE Document Recognition and Retrieval VI*, pp. 122-133, 1999

Choi, Y., Rasmussen, E.M. (2002) Users' relevance criteria in image retrieval in American history. *Information Processing and Management* 38(5): 695-726

Clough, P., Sanderson, M. (2003) The CLEF 2003 Cross Language Image Retrieval Task, in *Working Notes for the CLEF 2003 Workshop*, 21-22 August, Trondheim, Norway

Dunlop, M.D. & van Rijsbergen, C. J. (1993) Hypermedia and free text retrieval, *Information Processing and Management*, vol **29**(3).

Flank, S. (2002) Cross-Language Multimedia Information Retrieval, in the Proc. *6th Applied Natural Language Processing Conference*

Gollins, T. (2000) Dictionary Based Transitive Cross-Language Information Retrieval Using Lexical Triangulation, Masters Dissertation, Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP, Sheffield.

Harmandas, V, Sanderson, M., Dunlop, M.D. (1997) Image retrieval by hypertext links In the *Proceedings of the 20th ACM SIGIR conference*, Pages 296-303, 1997

Jones, G.J.F., New Challenges for Cross-Language Information Retrieval: Multimedia Data and the User experience. In Carol Peters (ed.). *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science 2069*, Springer 2001, pp 71-81.

Oard, D. (1997) Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries In *D-Lib Magazine*, http://www.dlib.org/

Oard, D., Gonzalo, J., Sanderson, M., López-Ostenero, F., Wang, J. (2004) Interactive Cross-Language Document Selection to appear in the journal of *Information Retrieval*

Paterson, C. (2002) The effectiveness of using machine translators to translate German and Portuguese into English when searching for images with English captions, Masters Dissertation, Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP, Sheffield.

Reid, N. (1999) The photographic collections in St. Andrews University Library. *Scottish Archives*, Vol. 5, 83-90

Resnik, P. (1997) Evaluating Multilingual Gisting of Web Pages, in *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval Electronic Working Notes*

Sormunen, E., Laaksonen, J., Keskustalo, H., Kekäläinen, J., Kemppainen, H., Laitinen, H., Pirkola, A., Järvelin, K., (1998) The IR Game - A Tool for Rapid Query Analysis in Cross-Language IR Experiments. *PRICAI '98 Workshop on Cross Language Issues in Artificial Intelligence*. pp. 22-32.

Voorhees, E. (1998): Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness, in *Proceedings of the 21st annual international ACM-SIGIR conference on Research and development in information retrieval*: 315-323