The Importance of Evaluation for Cross-Language System Development: the CLEF Experience

Carol Peters[1] **and Martin Braschler**[2]

[1]IEI-CNR, Area di Ricerca CNR, 56124 Pisa, Italy
carol@iei.pi.cnr.it
[2]Eurospider Information Technology, Zürich, Switzerland
martin.braschler@eurospider.com

Abstract

The aim of the Cross-Language Evaluation Forum (CLEF) is to develop and maintain an infrastructure for the evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes. Two CLEF evaluation campaigns have been held so far (CLEF 2000 and CLEF 2001); CLEF 2002 is now under way. The paper describes the objectives and the organisation of these campaigns, and gives a first assessment of the results. In conclusion, plans for future CLEF campaigns are reported.

Introduction

The last decade has seen a revolution in the way that information is disseminated and retrieved. The popularity of the Internet and the consequent global availability of networked information sources for an increasingly vast public have led to a strong demand for efficient cross-language information retrieval (CLIR) systems that allow users to search document collections in multiple languages and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages.

However, such systems are not easy to develop and work is generally still in an experimental stage. Approaches currently being tested imply the integration of tools and methodologies from the fields of information retrieval, natural language processing and human-computer interaction among others. A typical CLIR system will include components for (i) the matching of queries to documents over languages, (ii) the retrieval of information ranked according to relevance, and (iii) the presentation of results in a way that is easily interpreted by the user. Developers need to understand the contribution of each of these components to the overall effectiveness of their system. An intensive process of system testing and tuning is thus needed before the separate components can be implemented successfully in end-user applications.

The Text REtrieval Conference (TREC) series organized by the US National Institute of Standards and Technology (NIST) has demonstrated that system evaluation activities can have a very beneficial impact on the testing part of the system development life-cycle (Smeaton & Harman, 1997). For this reason, a track for cross-language system evaluation was organised at TREC for three years, from 1997-99, focusing on a small set of European languages. An overview can be found in Harman et al. (2001). At the end of 1999, it was decided to center the coordination of this activity in Europe, while TREC would move its attention to other language typologies.

The Cross-Language Evaluation Forum (CLEF) thus represents the continuation and expansion of the activity begun at TREC. CLEF[1] aims at promoting CLIR system development by providing the research community with an infrastructure for:

- testing and evaluation of information retrieval systems operating in both monolingual and cross-language contexts
- objective comparison of different systems and approaches
- exchange of experiences and know-how between R&D groups working in the field.

The design of the tasks offered by CLEF is studied to meet the needs of developers working mainly with European languages. However, strong links have also been forged with the other two major CLIR system evaluation activities: the TREC activity which is currently focusing on English/French to Arabic retrieval (Gey & Oard, 2001) and the NACSIS Test Collection for Information Retrieval (NTCIR) sponsored by the National Institute for Informatics of Tokyo which offers cross-language system evaluation for Asian languages (see Kando et al, 2001). The three initiatives (US, Asian and European) aim at creating a network of complementary activities in the cross-language system evaluation area.

In this paper, we describe the organization of the CLEF evaluation campaigns and list the main findings of

[1] CLEF 2000 and 2001 were sponsored by the DELOS Network of Excellence for Digital libraries; from October 2001, CLEF is funded by the European Commission under the IST programme (IST-2000-31002). The consortium members are: IEI-CNR, Pisa; IZ Sozialwissenschaften, Bonn; ELRA/ELDA, Paris; Eurospider, Zurich; UNED, Madrid; NIST, USA.

CLEF 2000 and 2001. The final section gives an idea of our plans and hopes for the future. For more details, the interested reader is referred to Peters & Braschler (2001).

The Methodology

Following the model used in TREC, CLEF uses a comparative evaluation approach and has adopted the well-known Cranfield methodology (Cleverdon, 1997): performance measures are calculated based on a test collection, sample queries and relevance assessments for these queries, with respect to the documents in the collection.

Following this philosophy and depending on the particular task to be performed and language(s) to be used, the effectiveness of information retrieval systems participating in the CLEF campaigns is evaluated as follows:

- the collection containing the appropriate test documents is indexed and inserted into the system
- the sample queries are indexed and run using the system against the document index
- the results are evaluated based on the relevance assessments.

In the following section, we describe the various tasks and test collections provided by CLEF and explain how the results of the participating systems are assessed and analysed.

The Tasks

CLEF provides a series of evaluation tracks designed to test different aspects of information retrieval system development. The intention is to encourage systems to move from monolingual searching to the implementation of a full multilingual retrieval service. The design of these tracks has been modified over the years in order to meet the needs of the research community. Here below we describe the tracks and tasks offered by CLEF 2002.

Multilingual Information Retrieval

This is the main task in CLEF. It requires searching a multilingual collection of documents for relevant items, using a selected query language. Multilingual information retrieval is a complex task, testing the capability of a system to handle a number of different languages simultaneously and to merge the results, ordering them according to relevance.

Bilingual Information Retrieval

In this track, any query language can be used to search just one of the CLEF target document collections. Many newcomers to CLIR system evaluation prefer to begin with the simpler bilingual track before moving on to tackle the more complex issues involved in truly multilingual retrieval.

Monolingual (non-English) IR

Until recently, most IR system evaluation focused on English. However, many of the issues involved in IR are language dependent. CLEF provides the opportunity for monolingual system testing and tuning, and for building test suites in other European languages but not English.

Mono- and Cross-Language Information Retrieval for Scientific Texts

The rationale for this task is to study CLIR on other types of collections, serving a different kind of information need. The information which is provided by domain-specific scientific documents is far more targeted than news stories and contains much terminology. It is claimed that the users of this type of collection are typically interested in the completeness of results. This means that they are generally not satisfied with finding just some relevant documents in a collection that may contain much more. Developers of domain-specific cross-language retrieval systems need to be able to tune their systems to meet this requirement. See Gey & Kluck (2001) for a discussion of this point.

For each of the tasks listed above, the participating systems construct their queries (automatically or manually) from a common set of statements of information needs (known as topics) and search for relevant documents in the collections provided, listing the results in a ranked list.

Interactive CLIR

The aim of the tracks listed above is to measure system performance mainly in terms of how good the document rankings are. However, this is not the only issue that interests the user. User satisfaction with an IR system will be based on a number of factors, depending on the functionality of the particular system. For example, the way in which the results of a search are presented is of great importance in CLIR systems where it is common to have users retrieving documents in languages which they do not understand. When users are unfamiliar with the target language, they need a presentation of the results which will permit them to easily and accurately select documents of interest, discarding others. An interactive track that focused on this document selection problem was experimented with success in CLEF 2001 (see Oard & Gonzalo, forthcoming).

The Test Collections

The main CLEF test collection is formed of sets of documents in different European languages but with common features (same genre and time period, comparable content); a single set of topics rendered in a number of languages; relevance judgments determining the set of relevant documents for each topic. A separate test collection is being created for systems tuned for domain-specific tasks.

Multilingual Corpus

The main document collection currently consists of nearly 1,000,000 documents in seven languages – Dutch, English, Finnish, French, German, Italian and Spanish. It contains both newswires and national newspapers. Spanish and Dutch were introduced for the first time in CLEF 2001 for different reasons. Spanish was included because of its status as the fourth most widely spoken language in the world. Dutch was added not only to meet the demands of the considerable number of Dutch participants in CLEF but also because it provides a challenge for those who want to test the adaptability of their systems to a new, less well-known language. Finnish has been included this year; its highly complex morphology and its membership of a different language family (Ugro-Finnic) with respect to the other European languages in the CLEF collection will provide an additional challenge for indexing.

Two distinct scientific collections are also available: the GIRT database of about 80,000 German social science documents, which has controlled vocabularies for English-German and German-Russian, and the Amaryllis multidisciplinary database of approximately 150,000 French bibliographic documents and a controlled vocabulary in English and French.

Topics

The participating groups derive their queries in their preferred language from a set of topics created to simulate user information needs. Following the TREC philosophy, each topic consists of three parts: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. The English version of a typical topic from CLEF 2001 is shown below:

> **Title:** U.N./US Invasion Haiti
>
> **Description:** Find documents on the invasion of Haiti by U.N./US soldiers.
>
> **Narrative:** Documents report both on the discussion about the decision of the U.N. to send US troops into Haiti and on the invasion itself. They also discuss the direct consequences.

The title contains the main keywords, the description is a “natural language” expression of the concept conveyed by the keywords, and the narrative adds additional syntax and semantics, stipulating the conditions for relevance assessment. The motivation behind these structured topics is to provide query “input” for all kinds of IR systems, ranging from simple keyword-based procedures to more sophisticated systems supporting morphological analyses, parsing, query expansion and so on. In the cross-language context, the transfer component must also be considered, whether dictionary or corpus-based, a fully-fledged MT system or other. Different query structures may be more appropriate for testing one or the other methodology.

For CLEF 2002, 50 such topics have been developed on the basis of the contents of the multilingual collection and topic sets have been produced in all seven document languages. Additional topic sets in Swedish, Portuguese Russian, Japanese, Chinese are now in preparation, other languages may be offered depending on demand. The same topic set is used for the multilingual, bilingual and monolingual tasks. Participants can thus choose to formulate their queries in any one of at least ten European or two Asian languages. Separate topic sets are developed for the scientific collections: in German, English and Russian for the GIRT task, and French and English for Amaryllis.

Relevance Judgments

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead, approximate recall figures are calculated by using pooling techniques. The results submitted by the participating groups are used to form a "pool" of documents for each topic and for each language by collecting the highly ranked documents from all the submissions. The assumption is that if a sufficient number of diverse systems contribute results to a pool, it is likely that a large percentage of all relevant documents will be included. All documents not included in the pool remain unjudged and are therefore assumed to be irrelevant. A main concern with such a pooling strategy is that if the number of not detected relevant documents is above a certain (low) threshold, the resulting test collection will be of limited future use in testing systems that did not contribute to the pool. A grossly incomplete pool would unfairly penalize such systems when calculating precision and recall measures. This pooling strategy was first adopted by TREC and has been subsequently employed by both NTCIR and CLEF. A number of studies have been made to test its validity (see Zobel, 1998; Voorhees, 2000).

A test of the completeness of the pools used for the CLEF 2000 campaign can be found in Braschler (2001). The test reveals that the completeness of the relevance assessments compares favorably to that of the assessments used for previous TREC ad-hoc campaigns. Relevance assessment of the documents in the pool is distributed over a number of different sites and performed in all cases by native speakers. The results are then analyzed centrally using recall and precision measures and run statistics are produced and distributed.

The problems involved in multilingual topic creation and relevance assessment are discussed in more detail in Kluck & Womser-Hacker (2002).

Results Analysis

The CLEF campaign evaluates all official submissions based on the relevance assessments. A variety of measures are calculated both for every individual submission and for overall statistics. The two central evaluation measures used are Recall and Precision. Recall measures the ability of a system to present all relevant items, whereas Precision measures the ability of the system to present only relevant items.

Recall $\rho_r(q) := \dfrac{\left|D_r^{rel}(q)\right|}{\left|D^{rel}(q)\right|}$

and

Precision $\pi_r(q) := \dfrac{\left|D_r^{rel}(q)\right|}{\left|D_r(q)\right|}$,

where $D_r(q) := \{d_1, ..., d_r\}$ is the answer set to query *q* containing the first *r* documents. The choice of *r* depends on the preference of the user: a low value for *r* implies that the user is interested in few, high-precision documents, whereas a high value for *r* means that the user conducts an exhaustive search. $D^{rel}(q)$ is the set of all relevant documents, and $D_r^{rel}(q) := D^{rel}(q) \cap D_r(q)$ is the set of relevant documents contained in the answer set (Schäuble 1997).

The two measures are somewhat in conflict: it is desirable in most cases to optimize for both measures, i.e. retrieving a maximum of relevant items while retrieving a minimum of irrelevant items, but systems that optimize for better recall often do so at the expense of precision, while systems that optimize for precision often adopt a conservative retrieval strategy that leads to lower recall. It is therefore important to analyze system performance in a variety of scenarios, such as precision at low recall levels, recall at low precision levels and balance between precision and recall. In CLEF, precision figures for a range of recall levels are published, as well as the popular average precision measure, which summarizes performance across various recall levels. Graphically, precision figures at multiple levels of recall can be visualized in the form of a recall precision graph (see Figure 1).

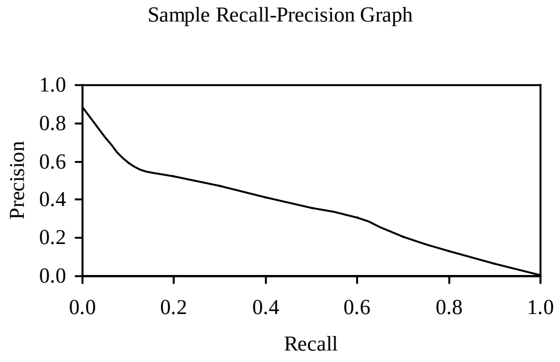

Figure 1: Sample CLEF Recall-Precision Graph

In the case of one particularly popular application for retrieval technology, the World Wide Web, recall is often seen as of secondary importance, since for most search requests there is an overwhelming number of potentially relevant hits. In such scenarios, high precision at low recall levels is increasingly popular as an evaluation metric. The CLEF policy of publishing a range of performance measures also caters for this application.

The goal of CLEF is a comparative evaluation of retrieval techniques. Absolute performance levels do not generally carry over across different experimental setups. CLEF facilitates result comparison by publishing both graphs that summarize overall results and by presenting a graphical comparison of individual results to median performance (see Figure 2).

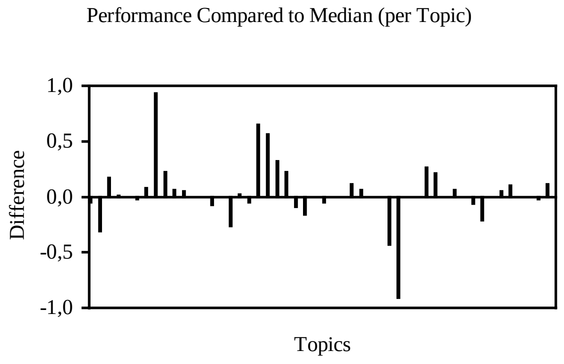

Figure 2: Comparison to Median by Topic

Starting with the 2001 campaign, CLEF also publishes an analysis of the statistical significance of performance differences observed between submissions by different participants (Braschler, forthcoming). Preliminary figures suggest that it is hard to achieve statistically significant performance differences, since the variability in performance between queries tends to be higher than the variability of performance between systems. Similar observations have been made before for TREC experiments (Tague-Sutcliffe & Blustein, 1995). CLEF tries to address this problem by producing topic sets that can be combined with earlier years' campaigns into larger sets, which helps to obtain more reliable figures for post-campaign evaluations. Additionally, CLEF also publishes average precision figures for individual experiments per topic, allowing comparison of systems for specific topics.

CLEF 2000 and 2001

The first two CLEF campaigns proved very successful. Over thirty groups from both academia and industry participated in CLEF 2001, up more than 50% with respect to the previous year. All the traditional approaches to CLIR system development were tried: machine translation, bilingual dictionary look-up, corpus-based approaches, conceptual networks, multilingual thesauri (for the domain-specific task). It was very interesting to witness the adjustments and refinements made to the basic strategies by many groups, and to see the results obtained.

A number of methods were adopted to index texts in multiple languages. The main diversification was between relatively simple stemming procedures and more complex morphological analysis. This appears to be very much a language dependent choice. Morphological processors were generally preferred for romance languages, whereas stemming appeared more popular for the Germanic languages. Both free and commercial stemmers were used, some groups attempted ad hoc simple generic "quick&dirty" stemming methods, and one group had considerable success with language independent indexing. The issue of decompounding in certain languages, such as Dutch and German was explored extensively. However, conflicting results were reported. While some groups reported substantial benefits, one group noted degradation of their retrieval results. A number of groups also adopted NLP strategies such as phrase identification and morphosyntactic analysis.

A major problem for CLIR systems is an insufficient coverage of the translation resource used. Many groups tackled this problem through an integration of different resources: MT, MRDs and corpora. Pivot language strategies were attempted in at least once case to translate from L1 -> L2 and N-gram based techniques were tried to match untranslatable words.

In 2001, a general trend noted was a move towards corpus-based statistical approaches. A number of groups experimented with automatically constructed resources of this type, and used them either for query translation or for resolving translation ambiguities (e.g. word sense disambiguation). Several groups used training data derived by mining the World Wide Web as input to their statistical models. Other groups used parallel or comparable corpora in several languages for a similar purpose. The statistical approaches were often combined with either machine-readable dictionaries or machine translation.

An evaluation activity of this type is important in that provides a forum in which traditional state-of-the-art methods can be compared against new approaches. One nice trend observed for the 2001 campaign was an increasing number of participants that tried techniques that were successfully introduced by other groups in previous campaigns. By integrating and modifying these techniques, new insights into their applicability are gained and new, more mature CLIR systems are produced. Stimulating this kind of cross-fertilization between research groups and system developers is a main goal of the CLEF project.

A final product of an evaluation campaign, or a series of campaigns, is a set of reusable test collections. The CLEF test-suite is a very valuable resource for developers testing and tuning their systems, but this is currently only available to registered participants. An objective of the CLEF project is to make the test-suites produced by the evaluation campaigns also accessible to the wider R&D community for benchmarking purposes.

Future Directions

Previous to the launching of the 2002 campaign, we conducted a survey in order to acquire input for the design of the tasks to be offered. Two types of users were considered: cross-language technology developers and cross-language technology deployers.

The first group was mainly represented by system developers who had previously participated in CLEF-campaigns or groups who had indicated interest in the CLEF evaluation activities. This group was already well aware of the objectives and potential of a system evaluation campaign and thus provided considerable useful input and concrete suggestions. The main recommendations made can be summed up in the following list:

- Increase the size and the number of languages in the multilingual test collection (both with respect to documents and topics);
- Provide the possibility to test on different text types (e.g. structured data);
- Provide more task variety (question-answering, web-style queries, text categorization);
- Study ways to test retrieval with multimedia data;
- Provide standard resources to permit objective comparison of individual system components (e.g. groups using a common retrieval system can compare the effect of their individual translation mechanisms);
- Focus more on user satisfaction issues (e.g. query formulation, results presentation).

The information acquired from the second questionnaire aimed at the system deployers was far less focussed,. However, three important points emerged clearly and invite reflection:

- Real-world applications do not just regard textual information;
- Document ranking is not the only factor of relevance to the end-user – ease-of-use, speed of response times and presentation of results in a comprehensible fashion are also high on the list of importance;
- Finally, there is a surprising lack of perception of the need for cross-language functionality, even in applications that are regularly handling information in multiple languages.

The first two points reinforce recommendations made by the system developers and encourage us to include tasks evaluating aspects that regard end-user satisfaction rather than system performance in isolation, and to consider media other than text, e.g. spoken document and/or image caption retrieval. The last one suggests that there is a strong need for more dissemination among technology deployers of the state-of-the-art of CLIR systems. Content and service providers should be made aware of the additional functionality that could be offered by their system with the inclusion of tools to handle multilingual information access.

As far as possible, the findings of this survey have been integrated into the definition of the CLEF 2002 campaign. Points that could not be taken up immediately will be considered for the future. As a first step, the size of the newspaper/newsagency collections and the number of languages covered have been increased. Language coverage in CLEF depends on two factors: the demand from potential participants and the existence of sufficient resources to handle the requirements of new languages.. Our goal is to be able to cover not only the major European languages but also some representative samples of minority languages, including members from each major group: e.g. Germanic, Romance, Slavic, and Ugro-Finnic languages. CLEF2002 has seen the addition of Finnish to the multilingual corpus, hopefully, 2003 will see the inclusion of a Russian collection. The topic languages this year should also include minority languages such as Basque and Catalan. Others will be considered in future years.

With respect to the demand for different types of texts and evaluation tasks, CLEF 2002 has seen the addition of the Amaryllis corpus to the multilingual collection of scientific documents; we now have a specific track dedicated to testing systems operating on different types of domain-specific collections. We are also considering the possibility of setting up a future track for text categorization task in multiple languages. We have had a number of contacts from potential participants and contacts are now under way with possible data providers.

In order to meet the demand regarding end-user related issues, the interactive track has been extended in 2002 and will be testing both user-assisted query translation and also document selection.

Last but certainly not least, as a first move towards handling multimedia, we are examining the feasibility of organising a spoken CLIR track in which systems would have to process and match spoken queries in more than one language against a spoken document collection. An experiment in this direction is being held this year within the framework of the DELOS Network of Excellence for Digital Libraries. The results will be presented at the annual CLEF Workshop in September 2002.

In conclusion, the results of the survey make it very clear that CLIR search functionality is perceived as just one component in a far more complex system cycle which goes from query formulation to results assimilation. In future years, we hope to go further in the extension and enhancement of CLEF evaluation tasks, moving gradually from a focus on cross-language text retrieval and the measuring of document rankings to the provision of a comprehensive set of tasks covering all major aspects of multilingual, multimedia system performance with particular attention to the needs of the end-user.

More information can be found on our Web site: http://www.clef-campaign.org/.

Acknowledgments

We gratefully acknowledge the support of all the data providers and copyright holders: The Los Angeles Times, for the English data collection; Le Monde S.A. and ELDA: European Language Resources Distribution Agency, for the French data; Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections; InformationsZentrum Sozialwissenschaften, Bonn, for the GIRT database; Institute de Information Scientifique et Technique, Vandoeuvre, France, for the Amaryllis database; Hypersystems Srl, Torino La Stampa, for the Italian data; Agencia EFE S.A. for the Spanish newswire; NRC Handelsblad; Algemeen Dagblad - PCM Landelijke dagbladen/Het Parool for the Dutch collections; Aamulehti Oyj for the Finnish newspaper data; Schweizerische Depeschenagentur (SDA) for French, Italian and German newswire data. Without their assistance, the CLEF evaluation activity would be impossible.

References

Braschler, M. (2001). CLEF 2000 – Overview of Results. In C. Peters (Ed.). Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069, Springer Verlag, pp 89-101

Braschler, M. (forthcoming). CLEF 2001 – Overview of Results. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.). Proceedings of CLEF 2001. Lecture Notes in Computer Science 2069, Springer Verlag, in print.

Cleverdon, C. (1997). The Cranfield Tests on Index Language Devices. In K. Sparck Jones and P. Willett (Eds.). Readings in Information Retrieval, pp 47-59. Morgan Kaufmann, 1997.

Gey, F.C. & Kluck, M. (2001). The Domain-Specific Task of CLEF – Specific Evaluation Strategies in Cross-Language Information Retrieval. In C. Peters (Ed.). Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069, Springer Verlag, pp 48-56.

Gey, F.C. & Oard, D.W. (2001). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries. NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001).

Harman, D., Braschler, M., Hess, M. Kluck, M., Peters, C., Schäuble, P. (2001). CLIR Evaluation at TREC. In C. Peters (Ed.). Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069, Springer Verlag, pp 7-23.

Kando, N., Aihara, K., Eguchi, K., Kato, H. (Eds.) (2001). Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, National Institute of Informatics (NII), ISBN 4-924600-89-X.

Kluck, M. & Womser-Hacker, C. (2002). Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. Paper to be published in LREC2002 Proceedings.

Oard, D.W. & Gonzalo, J. (forthcoming). The CLEF 2001 Interactive Track. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.). Proceedings of CLEF 2001. Lecture Notes in Computer Science 2069, Springer Verlag, in print.

Peters, C. & Braschler, M. (2001). Cross-Language System Evaluation: the CLEF Campaigns. Journal of the American Society for Information Science and Technology, 52(12), pp 1067-1072.

Schäuble, P.(1997). Content-Based Information Retrieval from Large Text and Audio Databases. Section 1.6 Evaluation Issues, Kluwer Academic Publishers, pp 22-29.

Smeaton A.F. & Harman, D. (1997). The TREC (IR) Experiments and their Impact on Europe, Journal of Information Science (23), pp 169-174.

Tague-Sutcliffe, J. & Blustein, J. (1995). A Statistical Analysis of the TREC-3 Data. In Proceedings of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-226. p385ff.

Voorhees, E.M. (2000). Variations in relevance judgments and the measure of retrieval effectiveness. Information Processing and Management (36), pp 697-716.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, pp 307-314.