

The CLEF 2001 Interactive Track

Douglas W. Oard and Julio Gonzalo

Human Computer Interaction Laboratory
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
oard@glue.umd.edu.edu,

WWW home page: <http://www.glue.umd.edu/~oard/>
and

Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
E.T.S.I Industriales, Ciudad Universitaria s/n, 28040 Madrid, SPAIN
julio@lsi.uned.es

WWW home page: <http://sensei.lsi.uned.es/~julio/>

Abstract. The problem of finding documents written in a language that the searcher cannot read is perhaps the most challenging application of cross-language information retrieval technology. In interactive applications, that task involves at least two steps: (1) the machine locates promising documents in a collection that is larger than the searcher could scan, and (2) the searcher recognizes documents relevant to their intended use from among those nominated by the machine. The goal of the 2001 Cross-Language Evaluation Forum's experimental interactive track was to explore the ability of present technology to support interactive relevance assessment. This paper describes the shared experiment design used at all three participating sites, summarizes preliminary results from the evaluation, and concludes with observations on lessons learned that can inform the design of subsequent evaluation campaigns.

1 Introduction

The problem of finding documents written in a language that the searcher cannot read is perhaps the most challenging application of Cross-Language Information Retrieval (CLIR) technology. In some cases (e.g., alerting the user to urgent new information), this might need to be a fully automatic process. In many applications, however, the effectiveness of fully automatic systems is limited by one or more of the following factors:

- The information need might initially be incompletely understood by the searcher.
- The information need might initially not be well articulated, either because the system's capabilities are underutilized or because the system's query language is insufficiently expressive.

- The ambiguity introduced by the use of natural (i.e., human) language within documents may cause the system to retrieve some documents that are not useful and/or to fail to retrieve some documents that are useful.

For this reason, automatic search technology is often embedded within interactive applications to achieve some degree of synergy between the machine’s ability to rapidly cull through enormous collections using relatively simple techniques and a human searcher’s ability to learn about their own information needs, to reformulate queries in ways that better express their needs and/or better match the system’s capabilities, and to accurately recognize useful documents within a set of a limited size (perhaps 10-100 documents). The goal of the experimental interactive track at the 2001 Cross-Language Evaluation Forum (which we call iCLEF) is to begin the process of exploring these issues in the context of cross-language information retrieval.

The process by which searchers interact with information systems to find documents has been extensively studied (for an excellent overview, see [1]). Essentially, there are two key points at which the searcher and the system interact: query formulation and document selection. Query formulation is a complex cognitive process in which searchers apply three kinds of knowledge—what they think they want, what they think the information system can do, and what they think the document collection being searched contains—to develop a query. The query formulation process is typically iterative, with searchers learning about the collection and the system, and often about what it is that they really wanted to know, by posing queries and examining retrieval results. Ultimately we must study the query formulation process in a cross-language retrieval environment if we are to design systems that effectively support real information seeking behaviors. We were concerned, however, that the open-ended nature of the query formulation process might make it difficult to agree on a sharp focus for quantitative evaluation in the near term. We therefore chose to focus on cross-language document selection for the initial iCLEF evaluation.

Interactive document selection is essentially a manual detection problem—given the documents that are nominated by the system as being of possible interest, the searcher must recognize which documents are truly of interest. The main Cross-Language Evaluation Forum (CLEF) track evaluates the effectiveness of systems that develop a ranked list of documents that are possibly (and hopefully!) relevant to a query, so we took that as our starting point. The searcher’s task thus becomes recognizing relevant documents in a language that they cannot read. Viewed from the perspective of system designers, the task is to present information (metadata, summaries, translations, etc.) that is sufficient to allow the user to make accurate relevance judgments.

Focusing on interactive CLIR is not actually be as a radical departure for CLEF as it might first appear. The principal CLEF evaluation measure—mean average precision (*MAP*)—actually models the automatic component of an interactive search process [2]. *MAP* is defined as:

$$MAP = E_i[E_j[\frac{j}{r(i,j)}]]$$

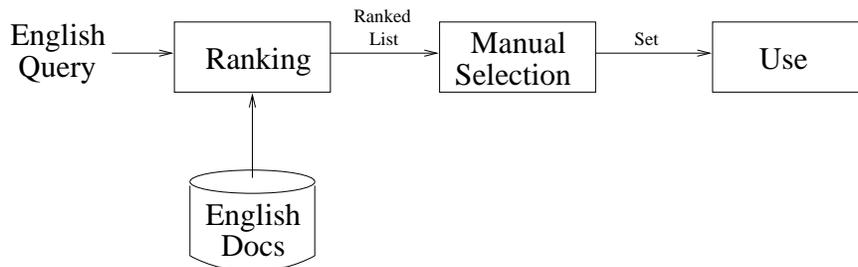


Fig. 1. A one-pass monolingual search process.

where $E_i[]$ is the sample expectation over a set of queries, $E_j[]$ is the sample expectation over the documents that are relevant to query i , and $r(i, j)$ is the rank of the j^{th} relevant document for query i . One way to think of MAP is as a measure of effectiveness for the one-pass interactive retrieval process shown in Figure 1 in which:

1. The searcher creates a query in a manner similar to those over which the outer expectation is computed.
2. The system computes a ranked list in a way that seeks to place the topically relevant documents as close to the top of the list as is possible, given the available evidence (query terms, document terms, embedded knowledge of language characteristics such as stemming, ...).
3. The searcher starts at the top of the list and examines each document (and/or summaries of those documents) until they are satisfied.
4. The searcher becomes satisfied after finding some number of relevant documents, but we have no *a priori* knowledge of how many relevant documents it will take to satisfy the searcher.
5. The searcher's degree of satisfaction is related to the number of documents that they need to examine before finding the desired number of relevant documents.

Implicit in this process is the assumption that the user can recognize relevant documents when they see them. It is that question that we sought to explore at iCLEF.

The remainder of this paper is organized as follows. Section 2 presents the basic experiment design that all three sites adopted and describes the shared evaluation resources that were provided. Section 3 then summarizes the research questions that each site explored and briefly summarizes some of the preliminary insights gained through cross-site comparison. Finally, Section 4 provides a preliminary recapitulation of some of the lessons we have learned that could inform the design of subsequent evaluation campaigns.

2 Experiment Design

Our experiment design closely follows the framework established over several years at the interactive track of the Text Retrieval Conferences (TREC).

2.1 Data

Document collection. We decided to use data from the CLEF 2000 campaign for several reasons:

- Ranked lists from existing automatic CLIR systems provide a representative sample of the input that an interactive document selection stage must be designed to handle.
- The use of a common set of frozen ranked lists enhanced the potential for cross-site comparisons.
- Relevance judgments for most of the top-ranked documents found in this way were already available, which made it possible for us to set the deadline for the interactive track about one month after the main CLEF 2000 task deadline in order to facilitate participation by teams that wished to participate in both tasks.
- Rights to use the CLEF 2000 collection for research purposes have already been arranged for CLEF participants.

We used top-1000 results from John Hopkins University for English documents (found for CLEF 2000 using French queries) and from the University of Maryland for French documents (found after CLEF 2000 using English queries) as the basis for forming the ranked lists that would be used in the experiments. We chose to support more than one document language because alternatives were needed in order to satisfy our requirement that teams recruit only searchers that were not familiar with the document language. These top-1000 results were then used to produce top-50 English and top-50 French results for each topic by first removing any document for which a relevance judgment was unavailable and then selecting the top 50 remaining documents. This process made it possible to use runs that had not been included in the original CLEF 2000 judging pools without the added complexity of scoring documents for which no CLEF relevance judgments were available.

As a baseline Machine Translation (MT) system, we chose Systran professional 3.0 because it is representative of state-of-the-art systems for language pairs in which there is considerable interest. Another factor favoring selection of Systran is that its use by popular freely-available Web page translation services makes it a *de facto* baseline for this task. We chose to translate the French documents into English and the English documents into Spanish for the baseline translations since those language pairs met the needs of teams that we knew were planning to participate. Use of the baseline translations was not required, so in principle it would have been possible for teams that preferred other language pairs to participate as well. In practice, all participating teams did choose to use the baseline translations for at least one of their two conditions.

Topics. For our experiment design we needed two “broad” topics that asked about some general subject that we thought would have many aspects, and two “narrow” topics that asked about some specific event. We selected those topics from among the 40 CLEF 2000 topics in the following manner:

- Discard topics that do not fall clearly into either the “broad” or the “narrow” category.
- Discard topics for which the relevance of a document could likely be judged simply by looking for a proper name (e.g. *Suicide of Pierre Berezovoy*).
- Favor topics that were relatively easy to judge for relevance based on:
 - a clear topic description, and
 - little need for specialized background knowledge.
- Favor topics with a greater number of known relevant documents in the top-50 for both languages.

Topic	Summary	Relevant Fraction	
		English	French
11 (broad)	<i>New constitution for South Africa</i>	36/50	27/50
13 (broad)	<i>Conference on birth control</i>	16/50	11/50
17 (narrow)	<i>Bush fire near Sydney</i>	6/50	2/50
29 (narrow)	<i>Nobel Prize for Economics</i>	2/50	3/50

Table 1. Selected topics

Our choice of topics according to these criteria turned out to be more limited than we had expected. Table 1 shows our choices and the density of relevant documents for each topic. One interesting outcome of our topic selection process is that it turned out that the narrow topics consistently had far fewer known relevant documents in the CLEF-2000 collection than the broad topics. Thus, for this collection, “narrow” roughly equates to “sparse” and “broad” roughly equates to “dense.” In addition to the topics chosen for the experiment, we suggested the use of topic 33 (*Cancer genetics*, a broad topic) for training searchers at the outset of their session. The same standard resources (top-50 lists and baseline translations) were therefore provided for topic 33 as well.

2.2 Search Procedure

The task assigned to each participant in an experiment was to begin at the top of a ranked list that had been produced by a cross-language retrieval system (see above) and to determine for as many documents in the list as practical in the allowed time whether that document was “relevant,” “somewhat relevant,” or “not relevant” to a topic described by a written topic description. The written

topic description included the text from the title, description, and narrative fields of the CLEF 2000 topic description. A maximum of 20 minutes was allowed for each topic, and participants were to be told that “more credit will be awarded for accurately assessing relevant documents than for the number of documents that are assessed, because in a real application you might need to pay for a high-quality translation [of] each selected document.” The participants were also afforded the ability to indicate if they were unsure of their assessment for a document, and they could also choose to leave some documents unassessed.

The participants were asked to complete eight questionnaires at specific points during their session:

- Before the experiment, about computer/searching experience and attitudes, and their degree of knowledge of the document collection, and their foreign language skills. (1)
- After assessing the documents with respect to each topic. (4)
- After completing the use of each system. (2)
- After the experiment, about system comparisons and to provide feedback on the experiment design. (1)

These questionnaires closely followed the design of the questionnaires used in recent TREC interactive track evaluations. The questionnaires that we used, among with additional forms for recording the experimenter’s observations during each search, can be found on the CLEF interactive track home page (which can be reached through <http://www.clef-campaign.org>). Each four-search session was designed to be completed in about three hours. This time included initial training, four 20-minute searches, all questionnaires, and two breaks (one following training, one between systems).

2.3 Presentation Order

We adopted a within-subject design in which each participant searched each topic with some system. Participants, topics and systems were distributed using a Latin square design in a manner similar to that used in the TREC interactive tracks. The presentation order for topics was varied systematically, with participants that saw the same topic-system combination seeing those topics in a different order. That design made it possible to control for fatigue and learning effects to some extent. An eight-participant presentation order matrix is shown in Table 2. The minimum number of participants was set at 4, in which case only the top half of the matrix would be used. Additional participants could be added in groups of 4, with the same matrix being reused as needed.

2.4 Evaluation

As our principal measure of effectiveness we selected an unbalanced version of van Rijsbergen’s F measure that we called F_α :

Participant	Block #1	Block #2
1	System 1: 11-17	System 2: 13-29
2	System 2: 11-17	System 1: 13-29
3	System 1: 17-11	System 2: 29-13
4	System 2: 17-11	System 1: 29-13
5	System 1: 11-17	System 2: 29-13
6	System 2: 11-17	System 1: 29-13
7	System 1: 17-11	System 2: 13-29
8	System 2: 17-11	System 1: 13-29

Table 2. Presentation order for topics and association of topics with systems.

$$F_{\alpha} = \frac{1}{\alpha/P + (1 - \alpha)/R}$$

where P is precision and R is recall [4]. Values of α above 0.5 emphasize precision, values below 0.5 emphasize recall. For this evaluation, $\alpha = 0.8$ was chosen, modeling the case in which missing some relevant documents would be less objectionable than finding too many documents that, after perhaps paying for professional translations, turn out not to be relevant. The CLEF relevance judgments are two-state (relevant or not relevant), so we treated all judgments other than “relevant” (“somewhat relevant,” “not relevant,” “not enough information”) as not relevant when computing F_{α} . For contrast, we computed $F_{0.2}$ (which modeled a recall-biased searcher) in addition to $F_{0.8}$, and participating teams were encouraged to explore additional measures that might better model cross-language retrieval tasks in which they were interested.¹

3 Results

We established an email reflector for teams that were interested in participating in the evaluation and other interested parties. Twenty people from 12 university, industry and government organizations joined that list. Three of those teams completed the experiment and submitted results: Universidad Nacional de Educación a Distancia (UNED) from Spain, the University of Maryland (UMD) from the USA, and the University of Sheffield (SHEF) from the United Kingdom. In this section we summarize the research questions explored by each team.

The **UNED** experiments used native Spanish speakers, Systran translations from English as a baseline, and “pseudo-translations” formed as lists of translated phrases as the contrastive condition. The hypotheses tested was that lists of translated phrases could permit faster judgments without significant loss in

¹ The more common subscript on F is β , which is defined such that $\alpha = 1/(\beta^2 + 1)$. $\alpha = 0.8$ corresponds to $\beta = 0.5$, which can be intuitively understood as valuing precision twice as much as recall. Similarly, $\alpha = 0.2$ corresponds to $\beta = 2.0$, which places a corresponding premium on recall.

precision. Eight monolingual Spanish-speaking searchers completed the task. In addition, a group of 8 searchers with a medium knowledge of English, and another 8-searcher group with a good knowledge of English, also completed the task.

The **University of Maryland** used four native English speakers to compare the utility of term-for-term gloss translations with results obtained using the baseline Systran translations. The hypothesis tested was that a combination of term-for-term gloss translation and query-term highlighting in the retrieved documents could provide a useful basis for relevance assessment.

The **University of Sheffield** used 8 native English-speaking searchers to compare monolingual and cross-language and document selection. The specific tasks included selecting French documents using Systran translations, and selecting documents from the (untranslated) English collection. Because both collections were used, the SHEF experiments offer a useful basis for comparison with both the UMD and UNED results.

Table 3 summarizes the results obtained for both languages. Figure 2 illustrates the French results using a recall-precision plot, and Figure 3 provides a similar depiction for English. For comparison, a naive searcher that marked every document as relevant would achieve a precision of 0.30 for English or 0.22 for French, with a recall of 1.0 in either case. This would have resulted in $F_{0.8} = 0.35$ for English and $F_{0.8} = 0.26$ for French.

English documents					French documents				
System	P	R	$F_{0.8}$	$F_{0.2}$	System	P	R	$F_{0.8}$	$F_{0.2}$
SHEF-Monolingual	.59	.40	.45	.39	UMD-MT	.76	.58	.61	.57
UNED-Phrases	.47	.34	.35	.32	SHEF-MT	.67	.46	.59	.48
UNED-MT	.48	.22	.28	.21	UMD-Gloss	.51	.27	.29	.26

Table 3. Overview of results.

The papers submitted by each team describe system-specific results. By looking across the three systems we are also able to make the following observations:

- The fact that every system achieved better precision than could have been obtained through the naive selection of every document suggests that every technique that was tried has some merit.
- The usefulness of Systran translations for this task appears to be consistent across sites (for French-to-English, at SHEF and UMD), but not across languages (where both precision and recall with English-to-Spanish translations were well below that achieved with French-to-English translations).
- Monolingual assessment appears to be substantially better (in both precision and recall) than cross-language assessment using Systran, and cross-language assessment using Systran appears to be substantially better (in both precision and recall) than the term-for-term gloss translation technique that was tried at UMD.

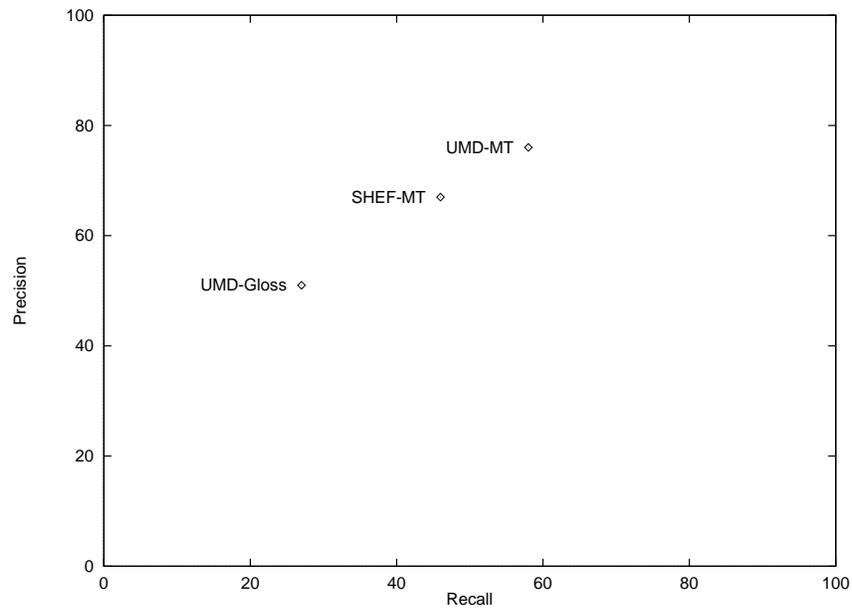


Fig. 2. Overview of French results.

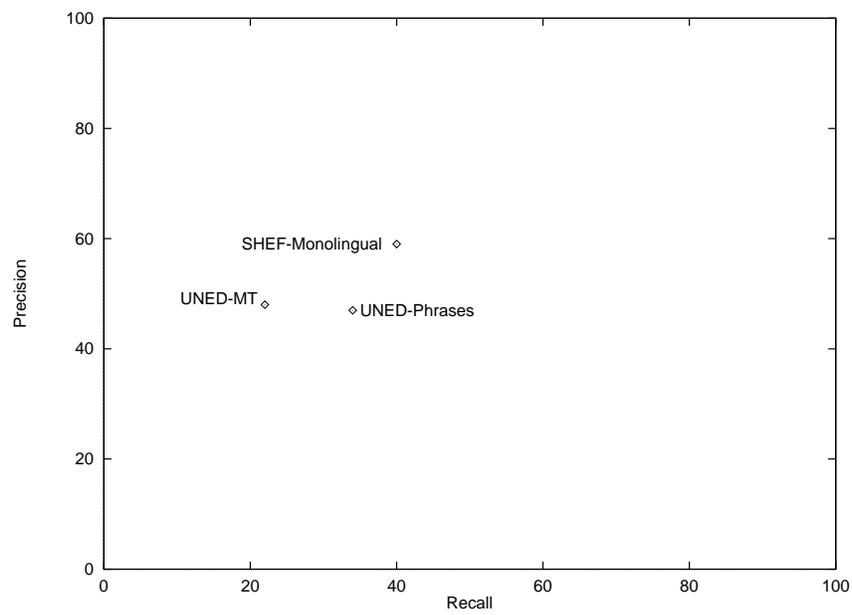


Fig. 3. Overview of English results.

- The display of translated phrases (UNED’s pseudo-translations) appears to increase recall with no adverse affect on precision.
- There was a very substantial difference between CLEF relevance judgments (which would receive a F measure of 1.0 for any α) and monolingual assessment at iCLEF.

There are several possible explanations for this last point:

- Any pair of assessors will naturally disagree about some judgments, and assessors that lack expertise in a topic typically exhibit less agreement than experts would.
- CLEF assessors must judge every document as *relevant/not relevant*, while our searchers could also choose *somewhat relevant, not enough information*, or leave the document unjudged.
- iCLEF searchers must make their judgments in a more sharply limited period.
- iCLEF searchers were given instructions that were intended to bias them in favor of precision. Pooled relevance assessment, by contract, places a premium on careful consideration of every document in the assessment pool.
- Assessors in a formal evaluation could discuss difficult judgments with other assessors, thereby reflecting some degree community consensus in those cases. The iCLEF searchers produce only personal opinions..
- CLEF assessors evaluate documents in an arbitrary order, while iCLEF searchers have additional information available (the order of the documents in the ranked list).

By characterizing the degree to which a time-constrained interactive searcher’s judgment might differ from that exercised to establish ground truth for an information retrieval evaluation, we have gained an unexpected insight that might prove useful in the design of adaptive filtering and relevance feedback evaluations, even in a monolingual context.

4 Looking to the Future

With three participating teams in this first year and the limited number of participants at each site, our conclusions must necessarily be preliminary. We have nonetheless learned a number of interesting things in these experiments. Our thinking on next steps is organized in two parts: what we might do to improve the evaluation of cross-language document selection, and how we might approach evaluation of some of the other tasks that are also important to interactive CLIR.

Some ideas that we are considering for future evaluations of document selection are:

- Consideration of measures other than F_α
- Establishing an agreed framework for statistical significance testing and then using that framework as a basis for establishing the minimum required number of participants in each experiment.

- Exploring experiment designs that could yield insight into the difference between monolingual and cross-language performance on the same document collection.
- Capturing separate values for confidence and relevance assessment, rather than treating “unsure” as an assessment value.
- Exploring tasks other than a simple yes/no decision (e.g., creating suitable ground truth for evaluating multi-valued relevance judgments, evaluating aspectual recall for topics that have a rich substructure, or designing a question answering task).
- Providing shared tools that can reduce barriers to participation in the evaluation campaign (e.g., user interface toolkits that include provisions for logging interactive relevance judgments).

There will be an iCLEF track in 2002, in which teams will have the opportunity to explore the effects of both interactive query translation and interactive document selection. The track design has not yet been finalized, but our intention is to measure the effect of interactive query translation as the difference in mean average precision obtained using automatic and interactive query translation techniques.

5 Conclusion

One of the most valuable products of iCLEF has been the emergence of a community of interest around the subject of interactive cross-language retrieval. One important part of this community of interest is a set of researchers that think of themselves as working on task-situated machine translation (where cross-language relevance assessment is the task). Task-based evaluation frameworks have recently been receiving greater attention from machine translation researchers (for example, see [3]). Addressing the CLIR challenge is naturally an interdisciplinary endeavor, and the potential for close links between CLIR and machine translation researchers should therefore be very much in our mutual interest.

Although only three sites participated in this first cooperative evaluation of interactive CLIR, we feel that we achieved our initial goals. We gained a better understanding of the issues that need to be addressed to conduct such evaluations, discovered other researchers with similar interests, and obtained some interesting results. We hope that our email reflector will help to nurture and grow that community as we discuss what we have learned and add people that will bring new perspectives. Next year’s iCLEF should therefore benefit in many ways from what we have learned. But regardless of what happens next year, we believe that iCLEF has been an example of CLEF at its best—discovering interesting questions and providing the resources needed to begin to answer them.

Acknowledgments

The authors are grateful to Carol Peters (CNR-IEI Pisa) for her support and encouragement, Paul Over (NIST) and Bill Hersh (OHSU) for generously offering advice and resources based on their experience in the TREC interactive track, Paul McNamee (Johns Hopkins APL) for providing top-1000 automatic English results, Gina Levow (Maryland) for providing top-1000 automatic French results, Clara Cabezas (Maryland) for producing the final document lists used in the evaluation, Jianqiang Wang (Maryland) for providing the Systran translations, and Fernando López-Ostenero (UNED) for managing the iCLEF Web page and email reflector, developing the evaluation scripts, and helping with many other aspects of the evaluation.

References

1. Marti A. Hearst. User interfaces and visualization. In Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors, *Modern Information Retrieval*, chapter 10. Addison Wesley, New York, 1999. <http://www.sims.berkeley.edu/~hearst/irbook/chapters/chap10.html>.
2. Douglas W. Oard. Evaluating interactive cross-language information retrieval: Document selection. In Carol Peters, editor, *Proceedings of the First Cross-Language Evaluation Forum*. 2001. To appear. <http://www.glue.umd.edu/~oard/research.html>.
3. Kathryn Taylor and John White. Predicting what MT is good for: User judgments and task performance. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Third Conference of the Association for Machine Translation in the Americas*, pages 364–373. Springer, October 1998. Lecture Notes in Artificial Intelligence 1529.
4. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.