A Language-Independent Approach to European Text Retrieval

Paul McNamee and James Mayfield
Johns Hopkins University Applied Physics Lab
11100 Johns Hopkins Road
Laurel, MD 20723-6099 USA
Paul.McNamee@jhuapl.edu
James.Mayfield@jhuapl.edu

We present an approach to multilingual information retrieval that does not depend on the existence of specific linguistic resources such as stemmers or thesaurii. Using the HAIRCUT system we participated in the monolingual, bilingual, and multilingual tasks of the CLEF-2000 evaluation. Our method, based on combining the benefits of words and character n-grams, was effective for both language-independent monolingual retrieval as well as for cross-language retrieval with translated queries. After describing our monolingual retrieval approach we compare a translation method using aligned parallel corpora to commercial machine translation software.

Background

The Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) is a research retrieval system developed at the Johns Hopkins University Applied Physics Lab (APL). One of the research areas that we want to investigate with HAIRCUT is the relative merit of different tokenization schemes. In particular we use both character n-grams and words as indexing terms. Our experiences in the TREC evaluations have led us to believe that while n-grams and words are comparable in retrieval performance, a combination of both techniques outperforms the use of a single approach. Through the CLEF-2000 evaluation we demonstrate that unsophisticated, language-independent techniques can form a credible approach to multilingual retrieval. We also compare query translation methods based on parallel corpora with automated machine translation.

Overview

We participated in the monolingual, bilingual, and multilingual tasks. For all three tasks we used the same 8 indices, a word and an n-gram based index in each of the four languages. Information about each index is provided in Table 1. In all of our experiments documents were indexed in their native language because we prefer query translation over document translation for reasons of efficiency.

	# docs	collection size (MB gzipped)	name	# terms	index size (MB)
English	110,282	163	enw	219,880	255
			en6	2,668,949	2102
French	44,013	62	frw	235,662	96
			fr6	1,765,656	769
German	153,694	153	gew	1,035,084	295
			ge6	3,440,316	2279
Italian	58,051	78	itw	278,631	130
			it6	1,650,037	1007

Table 1. Index statistics for the CLEF collection

We used two methods of translation in the bilingual and multilingual tasks. We used the Systran® translator to convert French and Spanish queries to English for our bilingual experiments and to convert English topics to French, German and Italian in the multilingual task. For the bilingual task we also used a method based on extracting translation equivalents from parallel corpora. Parallel English/French documents were most readily available to us, so we only applied this method when translating French to English.

Index Construction

Documents were processed using only the permitted tags specified in the workshop guidelines. First SGML macros were expanded to their appropriate character in the ISO-8859-1 character set. Then punctuation was eliminated, letters were downcased, and only the first two of a sequence of digits were preserved (e.g. 1920 became 19##). Diacritical marks were preserved. The result is a stream of blank separated words. When using n-grams we construct indexing terms from the same stream of words; the n-grams may span word boundaries but sentence boundaries are noted so that n-grams spanning sentence

boundaries are not recorded. Thus n-grams with leading, central, or trailing spaces are formed at word boundaries. We used 6-grams with success in the TREC-8 CLIR task [6] and decided to do the same thing this year. As can be seen from Table 1, the use of 6-grams as indexing terms increases both the size of the inverted file and the dictionary.

Query Processing

HAIRCUT performs rudimentary preprocessing on queries to remove stop structure, *e.g.*, affixes such as "… would be relevant" or "relevant documents should…." A list of about 1000 such English phrases was translated into French, German, and Italian using both Systran and the FreeTranslation.com translator. Other than this preprocessing, queries are parsed in the same fashion as documents in the collection.

The HAIRCUT HMM is a simple two-state model that captures both document and collection statistics [7]. After the query is parsed each term is weighted by the query term frequency and an initial retrieval is performed followed by a single round of relevance feedback. To perform relevance feedback we first retrieve the top 1000 documents. We use the top 20 documents for positive feedback and the bottom 75 documents for negative feedback, however we check to see that no duplicate or neo-duplicate documents are included in these sets. We then select terms for the expanded query based on three factors, a term's initial query term frequency (if any), the cube root of the (α=3, β=2, γ=2) Rocchio score, and a third term selection metric that incorporates an idf component. The top-scoring terms are then used as the revised query. After retrieval using this expanded and reweighted query, we have found a slight improvement by penalizing document scores for documents missing many query terms. We multiply document scores by a penalty factor:

$$PF = 1.0 - \left(\frac{\#\text{ of missing terms}}{\text{total number of terms in query}} \right)^{1.25}$$

We use only about one-fifth of the terms of the expanded query for this penalty function

	# Top Terms	# Penalty terms
words	60	12
6-grams	400	75

We conducted our work on a 4-node Sun Microsystems Ultra Enterprise 450 server. The workstation had 2 GB of physical memory and access to 50 GB of dedicated hard disk space.

The HAIRCUT system comprises approximately 25,000 lines of Java code.

Monolingual Experiments

Our approach to monolingual retrieval was to focus on language independent methods. We refrained from using language specific resources such as stoplists, lists of phrases, morphological stemmers, dictionaries, thesauri, decompounders, or semantic lexicons (e.g. Euro WordNet). We emphasize that this decision was made, not from a belief that these resources are ineffective, but because they are not universally available (or affordable) and not available in a standard format. Our processing for each language was identical in every regard and was based on a combination of evidence from word-based and 6-gram based runs. We elected to use all of the topic sections for our queries.

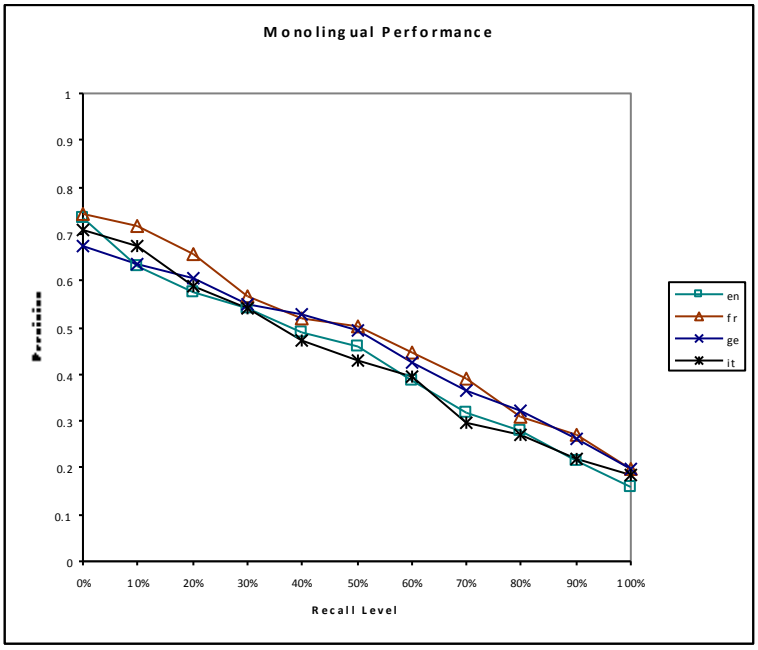

Figure 1. Precision-recall curves for the monolingual task. The English curve is unofficial and is produced from the bilingual relevance judgments.

The retrieval effectiveness of our monolingual runs is fairly similar for each of the four languages as evidenced by Figure 1. We expected to do somewhat worse on the Italian topics since the use of diacritical marks differed between the topic statements and the document collection; consistent with our 'language-independent' approach we did not correct for this. Given the generally high level of performance and the number of 'best' and 'above median' topics for the monolingual tasks (see Table 2), we believe that language independent techniques can be quite effective.

	avg prec	recall	# topics	# best	# ≥ median
aplmofr	0.4655	523 / 528	34	9	21
aplmoge	0.4501	816 / 821	37	10	32
aplmoit	0.4187	329 / 338	34	6	20

Table 2. Results for official monolingual submissions

One of our objectives was to compare the performance of the constituent word and n-gram runs that were combined for our official submissions. Figure 2 shows the precision-recall curves for the base and combined runs for each of the four languages. Our experience in the TREC-8 CLIR track led us to believe that n-grams and words are comparable, however each seems do perform slightly better in different languages. In particular, n-grams performed appreciably better on translated German queries, something we attribute to a lack of decompounding in our word-based runs. This trend was continued this year, with 6-grams performing just slightly better in Italian and French, somewhat better in German, but dramatically worse in our unofficial runs of English queries against the bilingual relevance judgments. We are stymied by the disparity between n-grams and words in English and have never seen such a dramatic difference in other test collectiions. Nonetheless, the general trend seems to indicate that combination of these two schemes has a positive effect as measured by average precision. Our method of combining two runs is to normalize scores for each topic in a run and then to merge multiple runs by the normalized scores.

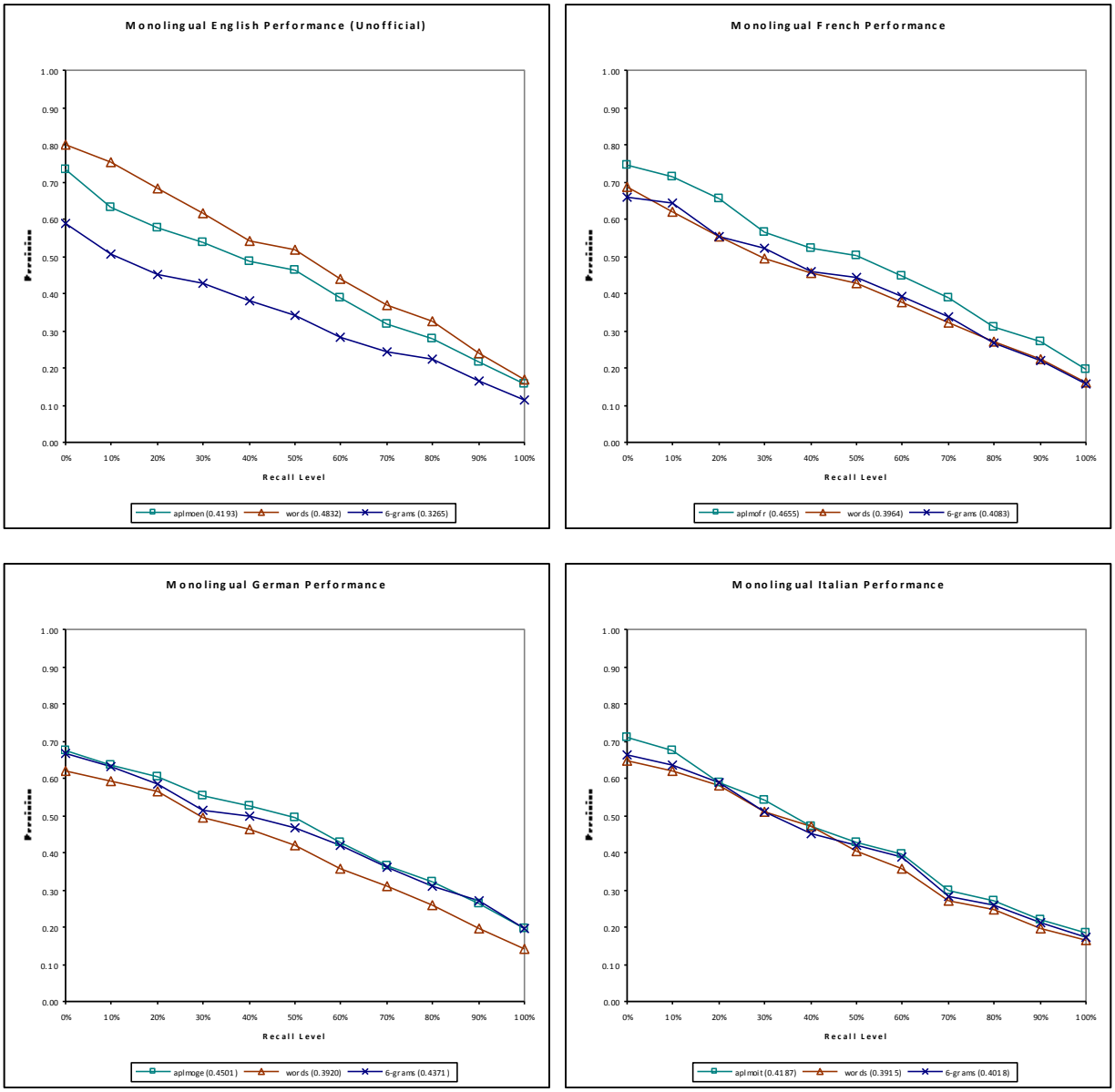

Figure 2. Comparison of retrieval performance using unstemmed words, 6-grams, and a combination of the two approaches for each of the four languages.

Bilingual Experiments

Our goal for the bilingual task was to evaluate two methods for translating queries, commercial machine translation software and a method based on aligned parallel corpora. While high quality MT products are available only for certain languages, the languages used most commonly in Western Europe are well represented. We used the Systran product which supports bi-directional conversion between English and the French, German, Italian, Spanish, and Portuguese languages. We did not use any of the domain specific dictionaries that are provided with the product.

The run, *aplbifrc*, was created by converting the French topic statements to English using Systran and searching the LA Times collection. As with the monolingual task both 6-grams and words were used separately and the independent results were combined. Our other official run using Systran was *aplbispa* that was based on the Spanish topic statements.

We only had access to large aligned parallel texts in English and French. We were therefore unable to conduct experiments in corpora-based translation in other languages. Our English / French dataset included text from the Hansard Set-A[5], Hansard Set-C[5], United Nations[5], RALI[8], and JOC[3] corpora. The Hansard data accounts for the vast majority of the collection.

	Description
Hansard Set-A	2.9 million aligned sentences

Hansard Set-C	aligned documents, converted to ~400,000 aligned sentences
United Nations	25,000 aligned documents
RALI	18,000 aligned documents
JOC	10,000 aligned sentences

Table 3. Description of the parallel collection used for *aplbifrb*

The process that we used for translating an individual topic is shown in Figure 3. First we perform a pre-translation expansion on a topic by running that topic in its source language on a contemporaneous expansion collection and extracting terms from top ranked documents. Thus for our French to English run we use the Le Monde collection to expand the original topic which is then represented as a weighted list of 60 words. Each of these words is then translated to the target language (English) using the statistics of the aligned parallel collection. We selected a single 'best' translation for each word and the translated word retained the weight assigned during topic expansion. Our method of producing translations is based on a term similarity measure similar to mutual information [2]; we do not use any dimension reduction techniques such as CL-LSI [4]. An example is shown for Topic C003 in Table 4. Finally we ran the translated query on the target collection in four ways, using both 6-grams and words and by using and not using relevance feedback.

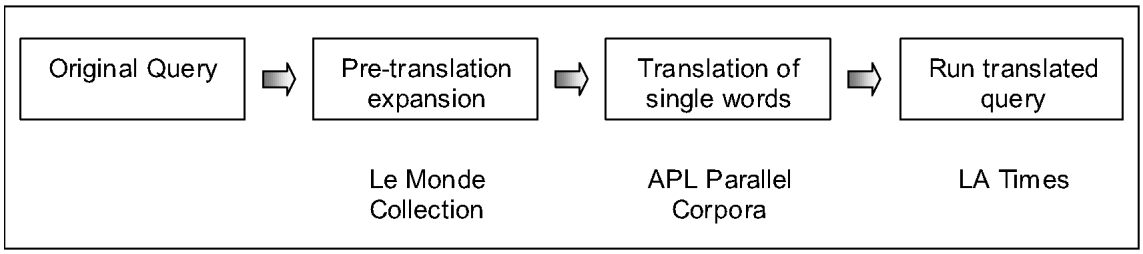

Figure 3. Processing steps for *aplbifrb*

Official French Query
<F-title> La drogue en Hollande

<F-desc> Quelle est la politique des Pays-Bas en matière de drogue?

<F-narr> Les documents pertinents exposent la réglementation et les décisions du gouvernement néerlandais concernant la vente et la consommation de drogues douces et dures.

Official English Query
<E-title>
Drugs in Holland

<E-desc>
What is the drugs policy in the Netherlands?

<E-narr> Relevant documents report regulations and decisions made by the Dutch government regarding the sale and consumption of hard and soft drugs.

Systran translation of French query
<F-title>
Drug in Holland

<F-desc>
Which is the policy of the Netherlands as regards drug?

<F-narr> The relevant documents expose the regulation and the decisions of Dutch government concerning the sale and the consumption of soft and hard drugs.

Figure 4. Topic C003 in the official French and English versions and as translated by Systran from French to English.

Weight	French	English	Weight	French	English
0.077601	drogue	drug	0.008583	prison	prison
0.068388	drogues	drugs	0.008490	suppression	removal
0.061828	douces	freshwater	0.008356	problème	problem
0.059526	dures	harsh	0.008344	produits	products
0.051063	consommation	consumer	0.008251	pénalisation	penalty
0.043705	matière	policy	0.008045	santé	health
0.040656	bas	low	0.007834	actuellement	now
0.037338	vente	sales	0.007831	consommateurs	consumers
0.035847	hollande	holland	0.007819	sévir	against
0.033313	néerlandais	netherlands	0.007743	réflexion	reflection
0.017426	cannabis	cannabis	0.007722	rapport	report
0.016166	stupéfiants	narcotic	0.007722	professeur	professor
0.015897	dépénalisation	decriminalization	0.007715	personnes	people
0.015011	usage	use	0.007714	souterraine	underground
0.014161	trafic	traffic	0.007706	partisans	supporters
0.013390	lutte	inflation	0.007678	sida	aids
0.013374	toxicomanie	drug	0.007667	débat	debate
0.012458	légalisation	legalization	0.007609	francis	francis
0.012303	héroïne	heroin	0.007578	europe	europe
0.011950	toxicomanes	drug	0.007561	membres	members
0.011725	usagers	users	0.009226	peines	penalties
0.010522	drogués	drug	0.009211	cocaïne	cocaine
0.010430	répression	repression	0.009106	alcool	alcohol
0.010379	prévention	prevention	0.008987	seringues	syringes
0.009892	loi	act	0.008926	risques	risks
0.009878	substances	substances	0.008829	substitution	substitution
0.009858	trafiquants	traffickers	0.008742	distinction	distinction
0.009813	haschich	hashish	0.008737	méthadone	methadone
0.009528	marijuana	marijuana	0.008721	dealers	dealers
0.009425	problèmes	problems	0.008698	soins	care

Table 4. Topic C003. French terms produced during pre-translation expansion and single word translation equivalents in English derived from parallel texts.

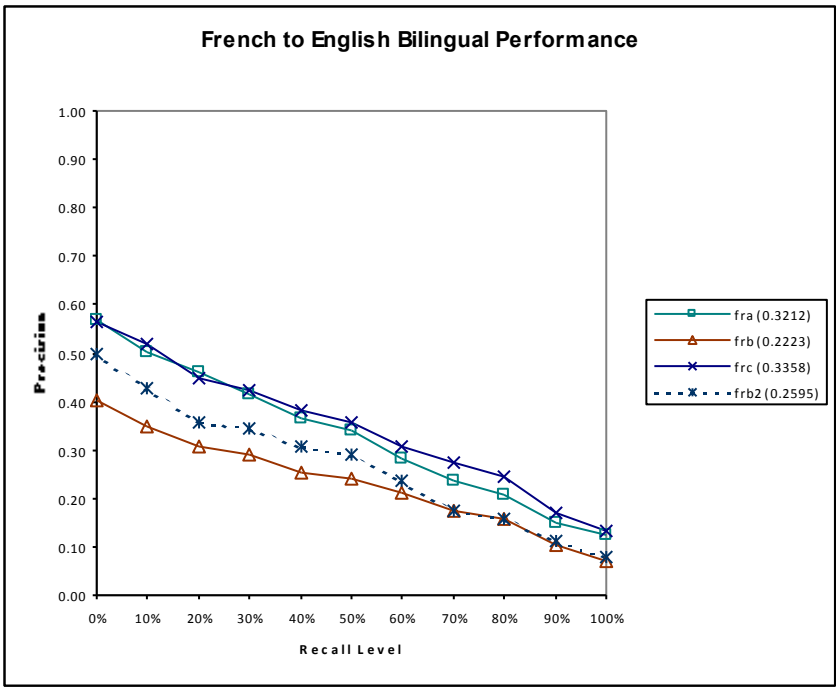

Figure 5. Comparison of *aplbifra* (combination), *aplbifrb* (parallel corpora), and *aplbifrc* (Systran).

We obtained superior results using translation software instead of our corpora-based translation. The precision-recall graph in Figure 5 shows a clear separation between the Systran-only run (*aplbifrc*) with average precision 0.3358 and the corpora-only run (*aplbifrb*) with average precision of 0.2223. We do not interpret this difference as a condemnation of our approach to corpus-based translation. Instead we agree with Brachler et al. that "MT cannot be the only solution to CLIR [1]." Both translation systems and corpus-based methods have their weaknesses. A translation system is particularly susceptible to named entities not being found in its dictionary. Perhaps as few as 3 out of the 40 topics in the test set mention obscure names: topics 2, 8, and 12. Topics 2 and 8 have no relevant English documents, so it is difficult to assess whether the corpora-based approach would outperform the use of dictionaries or translation tools on these topics. The run *aplbifra* is simply a combination of *aplbifrb* and *aplbifrc* that we had expected to outperform the individual runs.

There are several reasons why our translation scheme might be prone to error. First of all, the collection is largely based on the Hansard data, which are transcripts of Canadian parliamentary proceedings. The fact that the domain of discourse in the parallel collection is narrow compared to the queries could account for some difficulties. And the English recorded in the Hansard data is formal, spoken, and uses Canadian spellings whereas the English document collection in the tasks is informal, written, and published in the United States. It should be also noted that generating 6-grams from a list of words rather than from prose leaves out any n-grams that span word boundaries; such n-grams might capture phrasal information and be of particular value. Finally we had no opportunity to test our approach prior to submitting our results; we are confident that this technique can be improved.

With some post-hoc analysis we found one way to improve the quality of our corpus-based runs. We had run the translated queries both with, and without the use of relevance feedback. It appears that the relevance feedback runs perform worse than those without this normally beneficial technique. The dashed curve in Figure 5 labeled 'frb2' is the curve produced when relevance feedback is not used with the corpora-translated query. Perhaps the use of both pre-translation and post-translation expansions introduces too much ambiguity about the query.

Below are our results for the bilingual task. There were no relevant English documents for topics 2, 6, 8, 23, 25, 27, and 35, leaving just 33 topics in the task.

	avg prec	recall	# best	# ≥ median	method
aplbifra	0.3212	527 / 579	6	27	Combine aplbifrb/aplbifrc
aplbifrb	0.2223	479 / 579	4	23	Corpora FR to EN
aplbifrc	0.3358	521 / 579	7	23	Systran FR to EN
aplbispa	0.2595	525 / 579	5	27	Systran SP to EN

Table 5. Results for official bilingual submissions

Multilingual Experiments

We did not focuse our efforts on the multilingual task. We selected English as the topic language for the task and used Systran to produce translations in French, German, and Italian. We performed retrieval using 6-grams and words and then performed a multi-way merge using two different approaches, merging normalized scores and merging runs by rank.

	avg prec	recall	# best	# ≥ median	method
aplmua	0.2391	1698 / 2266	1	30	rank
aplmub	0.1924	1353 / 2266	3	23	score

Table 6. Results for official multilingual submissions

The large number of topics with no relevant documents in the collections of various languages suggests that the workshop organizers were successful in selecting challenging queries for merging. It seems clear that more sophisticated methods of multilingual merging are required to avoid a large drop in precision from the monolingual and bilingual tasks.

Conclusions

The CLEF workshop provides an excellent opportunity to explore the practical issues involved in cross-language information retrieval. We approached the monolingual task believing that it is possible to achieve good retrieval performance using language-independent methods. This methodology appears to have borne out based on our results using a combination of words and n-grams. For the bilingual task we kept our philosophy of simple methods, but also used a high-powered machine translation product. While our initial efforts using parallel corpora were not as effective as those with machine translated queries, the results were still quite credible and we are confident this technique can be improved further.

References

[1] M. Brachler, M-Y. Kan, and P. Schauble, 'The SPIDER Retrieval System and the TREC-8 Cross-Language Track.' In E. M. Voorhees and D. K. Harman, eds., Proceedings of the E*ighth Text REtrieval Conference (TREC-8)*. To appear.

[2] K. W. Church and P. Hanks, 'Word Association Norms, Mutual Information, and Lexicography.' In *Computational Linguistics*, 6(1), 22-29, 1990.

[3] European Language Resource Association (ELRA), http://www.icp.grenet.fr/ELRA/home.html

[4] T. K. Landauer and M. L. Littman, 'Fully automated cross-language document retrieval using latent semantic indexing.' In the *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*. 31-38, 1990.

[5] Linguistic Data Consortium (LDC), http://www.ldc.upenn.edu/

[6] J. Mayfield, P. McNamee, and C. Piatko, 'The JHU/APL HAIRCUT System at TREC-8.' In E. M. Voorhees and D. K. Harman, eds., *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. To appear.

[7] D. R. H. Miller, T. Leek, and R. M. Schwartz, 'A Hidden Markov Model Information Retrieval System.' In the Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR-99), pp. 214-221, August 1999.

[8] Recherche Appliquée en Linguistic Informatique (RALI), http://www-rali.iro.umontreal.ca/