Cross-Language Information Retrieval: Experiments Based on CLEF 2000 Corpora

Jacques Savoy
Institut interfacultaire d'informatique
Université de Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

Abstract

Search engines play an essential role in the usability of Internet-based information systems and without them the Web would be much less accessible, and at the very least would develop at a much slower rate. Given that non-English users now tend to make up the majority in this environment, our main objective is to analyze and evaluate the retrieval effectiveness of various indexing and search strategies based on test-collections written in four different languages: English, French, German, and Italian. Our second objective is to describe and evaluate various approaches that might be implemented in order to effectively access document collections written in another language. As a third objective, we will explore the underlying problems involved in searching document collections written in the four different languages, and we will suggest and evaluate different database merging strategies capable of providing the user with a single unique result list.

Keywords: Cross-language information retrieval; bilingual information retrieval; French, German, Italian languages; database merging strategies; evaluation.

1. Introduction

The increasing amount of information available on the Web means new and challenging problems are being confronted by the information retrieval community, one being the need for effective access to documents not written in the English language. There is currently an exponential growth in the amount of information available on the Web, with increasing amounts of resources being written in languages other than English. Based on recent statistics (February 2001), out of 313 million available pages on the Web, 68.4% of the pages are written in English (EMarketer, 2001; see also http://www.netsizer.com). Japanese (5.85%) is the second most popular language, followed by German (5.77%), Chinese (3.87%)

and French (2.96%). However, according to Global Reach (2001), for the majority of Internet users, English is not their first language. In 2001, those accessing the Internet in English accounted for 43%, compared to 9.2% for both Chinese and Japanese users. These numbers are difficult to estimate however due to the fact that some users may use other languages. For example, an estimated 32 million Americans will switch from English to another language to access the Web at home (mostly Spanish in this case), while 31.8% of users access it in one of the European languages (excluding English, mostly in Spanish, German, Italian and French). Moreover, these proportions seem to be changing quite rapidly; for example, in 1996 the English online population represented 80% of the total while in 2005 is has been estimated that this proportion will be only 29% (Global Reach, 2001). Thus, there is a real need to promote retrieval systems that can provide access to information without encountering language or cultural barriers.

In addition to the Web, various IR systems access documents in other contexts including digital libraries, newspapers, government archives and records, as well as legal and court decision documentation. In all cases, given the increasing volume of documents written in languages other than English and we must design and evaluate IR systems able to effectively access those documents and collections written in languages other than English.

For this reason, developing monolingual search engines would not be the more adequate or appropriate solution. For example, in multilingual countries such as Switzerland, the Federal Supreme Court may have to document legal cases, or parts of them, in German, French or Italian without providing translations into the other official languages. Similarly, Canada's Supreme Court has to write and document its decisions in either English or French. Also worth considering are the books and documents available in various languages in our libraries, in multinational companies or large international organizations (e.g., World Trade Organization, European Parliament or commissions), where the typical user needs to overcome various language barriers. For example, they may write a request in one language and yet wish to retrieve documents written in other languages. While some may need information written in various languages, they can usually read documents in other languages but cannot formulate a query in those language or, at least cannot provide reliable search terms to retrieve the documents being searched (Oard and Resnik, 1999). In other circumstances, monolingual users may want to retrieve documents in another language and then automatically or manually translate the texts retrieved into their own language. Finally, there are many documents in other languages containing information in non-textual formats

such as images, graphics and statistics could be made accessible to monolingual users, based on requests written in a different language.

Thus there is a real need to promote multilingual retrieval, and the Cross-Language Evaluation Forum (CLEF (Peters, 2001)) coordinated by the DELOS Network of Excellence on Digital Libraries (http://www.ercim.org/delos/), was founded to study and evaluate various multiple language information access technologies. One of its goals was to develop various non-English test-collections, some of which will be used in this paper.

This paper is organized as follows: Chapter 2 describes monolingual information retrieval systems dealing with document collections written in English, French, Italian and German. Chapter 3 illustrates and evaluates various approaches used to resolve bilingual information retrieval problems. In this case, the English set of requests provided in the CLEF 2000 test suite are translated into another language and the search is done on documents written in this target language. In Chapter 4 we investigate the underlying problems of searching corpora containing documents written in English, French, German and Italian, based on requests written only in English.

2. Evaluation of various indexing and searching strategies

Most European languages (including French, Italian, German) share many of the same characteristics with the language of Shakespeare (e.g., word boundaries marked in a conventional manner, variant word forms generated by adding an affix to the stem form of a lexeme, etc.). Any adaptation of indexing or search strategies thus means the creation of general stopword lists and fast stemming procedures. Stopword lists contain non-significant words that are removed from a document or a search request before the indexing process is begun. Stemming procedures try to remove inflectional and derivational suffixes in order to conflate word variants into the same stem or root. In attempting to resolve this problem for the French language, for example, it is important to remember that most European languages involve more complex morphologies than does the English language (Sproat, 1992).

This chapter will deal with some of these issues, and is organized as follows: Section 2.1 contains an overview of our four test-collections, Section 2.2 describes our general approach to building stopword lists and stemmers for use with languages other than English. This section also demonstrates how working with multi-lingual documents is more complex than just correctly handling the diacritical characters not usually present in English

collections (with some exceptions, such as "à la carte" or "résumé"). Section 2.3 depicts the various vector space term weighting schemes used in this paper together with the Okapi probabilistic model. Section 2.4 evaluates these search models using four test-collections and queries written in English, French, Italian and German. Finally, Section 2.5 describes the indexing scheme we use for German collections, based on 5-grams instead of words.

2.1. Overview of the test-collections

One of the main outcomes of the first CLEF workshop was the creation of various test-collections that would be available in four different languages. Within these corpora are included from 33 to 40 topics (see Appendix 2), and the requests reflect a diversity of information needs (such as "architecture in Berlin", "drug use and soccer", "solar temple" or "privatisation of German rail") rather than being limited to a narrow subject range. Following the TREC model, each topic was structured in three logical sections, namely a brief title, a one-sentence description and a narrative part specifying the relevance assessment criteria (see Table 1). For CLEF 2000, four groups participated in the construction of the topic sets and the original topics were composed in four languages (around 10 topics per collection). Subsequently these selected topics were manually translated in order to produce four complete sets of topics in English, French, Italian and German, as shown in Table 1. In order to provide complete topic coverage, this final set included "local topics" (having the explicit goal of hitting only one or two collections or languages) within both national and international interest areas (which might extract document in all languages).

Given that most Web queries were relatively short (2.21 terms according to Jansen *et al.* (2000) or 2.16 terms according to Spink *et al.* (2001)), our experiments were mainly based on information contained in the Title section, and our requests had an average length of 2.75 indexing terms. This limited query size also reflects the situation experienced by users whose level of foreign language knowledge would not permit the effective formulation of long requests, and who would thus tend to write relatively short queries. Of course in other environments, the length of queries submitted may be longer (e.g., in commercial IR systems, a mean query length of 14.8 search terms was reported by Spink and Saracevic (1997)). To reflect this second type of user requirement, we also studied queries based on the content of the Descriptive and Narrative logical sections.

The corpora used in our experiments consisted of national newspapers such as the *Los Angeles Times*, *Le Monde* (French), *La Stampa* (Italian), and *Der Spiegel* and *Frankfurter*

Rundschau (German). They are similar in content and subject matter (general news) and extracted during the same year (1994). As shown in Table 2, these corpora are of various sizes, with the English and German collections being twice the volume of the French and Italian sources. On the other hand, the mean number of distinct indexing terms per document is relatively similar across the corpora (around 180), and this number is a little bit smaller for the German corpus (145.66).

<num> C001
<E-title> Architecture in Berlin
<E-desc> Find documents on architecture in Berlin.
<E-narr> Relevant documents report, in general, on the architectural features of Berlin or, in particular, on the reconstruction of some parts of the city after the fall of the Wall.
<F-title> Architecture à Berlin
<F-desc> Trouver des documents au sujet de l'architecture à Berlin.
<F-narr> Les documents pertinents parlent, en général, des caractéristiques architecturales de Berlin ou, en particulier, de la reconstruction de certaines parties de cette ville après la chute du mur.
<I-title> Architettura a Berlino
<I-desc> Trova documenti che riguardano l'architettura a Berlino.
<I-narr> I documenti rilevanti parlano, in generale, degli aspetti architettonici di Berlino o, in particolare, della ricostruzione di alcuni parti della città dopo la caduta del Muro.
<G-title> Architektur in Berlin
<G-desc> Dokumente über Architektur in Berlin sind gesucht.
<G-narr> Relevante Dokumente berichten entweder allgemein über die Architektur in Berlin oder speziell über die Rekonstruktion von Teilen der Stadt nach dem Fall der Mauer.

Table 1: Sample CLEF 2000 topic statement showing all languages

The data in Table 2 also illustrates how the mean number of relevant documents per request can vary across the corpora. The German collection resulted in the greatest number of relevant items (821) and the highest mean per query (22.19). The Italian corpus on the other hand contained fewer relevant papers (338) and with a lower mean number of pertinent items per request (9.94).

During the indexing process, we retained only the following logical sections: <TITLE>, <HEADLINE>, <TEXT>, <LEAD> and <LEAD1> from the original documents, and we ignored other logical sections (even sections containing manually assigned index terms). From topic descriptions, we automatically removed certain phrases such as "Relevant document report …", "Find documents that give …", "Trouver des documents qui parlent …", "Sono valide le discussioni e le decisioni …" or "Relevante Dokumente berichten …".

	English	French	Italian	German
Size (in MB)	425 MB	157 MB	193 MB	383 MB
# of documents	113,005	44,013	58,051	153,694
# of distinct index terms / document				
mean	167.33	189.34	182.23	145.66
standard deviation	126.32	152.07	112.23	137.18
median	138	147	177	110
maximum	1,812	1,724	1,405	2,593
minimum	2	4	12	1
max df	69,082	40,054	55,690	57,711
# of indexing terms / document				
mean	273.85	285.60	239.31	183.16
standard deviation	246.88	236.13	156.06	198.17
median	212	216	230	132
maximum	6,087	3,963	3,800	6,642
minimum	2	12	13	1
# of queries	33	34	34	37
# of relevant documents	579	528	338	821
# of distinct relevant documents	545	507	336	816
mean relevant document per request	17.55	15.53	9.94	22.19
standard deviation	13.38	14.00	9.87	23.88
median	15	12	6.5	14
maximum	51 (#q:11)	62 (#q:5)	42 (#q:7)	101 (#q:5)
minimum	1 (#q:4)	1 (#q:22)	1 (#q:21)	1 (#q:6)

Table 2: Test-collection statistics

2.2. Stopword lists and stemming procedures

We defined a general stopword list made up of many words determined to be of no use during retrieval, but very frequently found in document content. These stopword lists were developed for two main reasons: Firstly, we hoped that each match between a query and a document would be based only on pertinent indexing terms. Thus, retrieving a document just because it contains words like "be", "your" and "the" in the corresponding request does not constitute an intelligent search strategy. These non-significant words represent noise, and may actually damage retrieval performance because they do not discriminate between relevant and irrelevant documents. Secondly, by using them we could reduce the size of the inverted file, hopefully in the range of 30% to 50%.

English and French stopword lists were already available (Fox, 1990), (Savoy, 1999). For German and Italian, we established a general stopword list by following the guidelines described in (Fox, 1990). Firstly, we sorted all word forms appearing in our corpora according to their frequency of occurrence and extracted the 200 most frequently occurring

words. Secondly, we inspected these lists in order to remove all numbers (e.g., "1994", "1"), plus all nouns and adjectives more or less directly related to the main subjects of the underlying collections. For example, the German word "Prozent" (ranking 69) or the Italian noun "Italia" (ranking 87) were removed from the final list; since from our point of view, such words are only useful as indexing terms in certain circumstances. Thirdly, we included certain words that contain no information, even though they did not appear in the first 200 most frequent words. For example, we added various personal or possessive pronouns (such as "meine" ("my" in German), prepositions ("nello" ("in the" in Italian)) and conjunctions ("où" ("where" in French)). The presence of homographs represents another debatable issue, and to some extent, we had to make arbitrary decisions concerning their inclusion in stopword lists. For example, the French word "son" can be translated as "sound" or "his", and the French term "or" as "thus/therefore" or "gold".

The resulting stopword list thus contained a large number of pronouns, articles, prepositions and conjunctions. As in various English stopword lists, there were also some verbal forms ("sein" ("to be" in German), "essere" ("to be" in Italian), "sono" ("I am" in Italian)). For our experiments we also used the stoplist provided by the SMART system (571 English words), 217 French words, 431 Italian words, 294 German words (these stopword lists are available at http://www.unine.ch/info/clef/).

After removing high frequency words, an indexing procedure tries to conflate word variants into the same stem or root using a stemming algorithm. In developing this procedure for the French, Italian and German languages, it is important to remember that these languages have more rich and complex morphologies than does the English language (Sproat, 1992). As a first approach, we intended to remove only inflectional suffixes so that singular and plural word forms or feminine and masculine variants conflate to the same root. We think that indexing verbs for Italian, French or German is not of primary importance, as compared to nouns and adjectives which tend to be more semantically significant. Moreover, in a previous study (Savoy, 1999), we have tried various more complex stemmers for the French language without obtaining any improvement in retrieval effectiveness. However more sophisticated schemes have already been proposed for the removal of derivational suffixes (e.g., «-ize», «-ably», «-ship» in the English language), such as the stemmer developed by Lovins (1968) based on a list of over 260 suffixes, while that of Porter (1980) looks for about 60 suffixes.

A "quick and dirty" stemming procedure has already been developed for the French language (Savoy, 1999) that attempts to remove frequent inflectional suffixes from nouns and adjectives. Based on these same ideas, we implemented a stemming algorithm for the Italian and German languages (the C code for these stemmers can be found at http://www.unine.ch/info/clef/). In Italian, the main inflectional rule is to modify the final character (e.g., «-o», «-a» or «-e») into another (e.g., «-i», «-e»). As a second rule, Italian morphology can also alter the final two letters (e.g., «-io» in «-o», «-co» in «-chi», «-ga» in «-ghe»). In German, a few rules can be applied to obtain the plural form of words (e.g., "Frau" into "Frauen" (woman), "Bild" into "Bilder" (picture), "Sohn" into "Söhne" (son), "Apfel" into "Äpfel" (apple)), but the suggested algorithms do not account for person and tense variations, nor the morphological variations found in verbs.

Finally, most European languages manifest other morphological aspects not taken into account by our approach, with compound word constructions being just one example (i.e., concatenation of two or more lexeme's stems to form another word, e.g., handgun, worldwide). In a similar vein, Italian plural forms may alter letters within a word, for example "capoufficio" (chief secretary) becomes "capiufficio" in its plural form, yet the modification of the stem "capo" does not follow a general rule (e.g., "capogiro" gives "capogiri" (dizziness) in its plural form). In German and in most Germanic languages compound words are widely used and they cause more difficulties than does English. For example, a life insurance company employee would be "Lebensversicherungsgesellschaftsangestellter" (Leben + S + Versicherung + S + Gesellschaft +S + Angestellter for life + insurance + company + employee). The augment (i.e. the letter «S» in our previous example) is not always present (e.g., "Bankangestelltenlohn" built as Bank + Angestellten + Lohn (salary)). Since compound construction is so widely used, it is almost impossible to build a German dictionary providing quasi-total coverage of this language. This linguistic feature also complicates the automatic translation of German documents. For example, in the word "Wagenleiter" (combining Wagen+Leiter meaning "car+ladder"), we emphasize the second word (e.g., to describe the ladder of a fire engine) while the German word "Leiterwagen" (chart) reflects the fact that the chart is made up of ladder-like components.

This concatenation process is also found in other languages. In Turkish, for example, the phrase "çöplüklerimizdekilerdenmiydi" combines çöp+lük+ler+imiz+de+ki+ler+den+mi+y+di to mean "was it from those that were in our

garbage cans?". However, this kind of long construction is more a linguistic game in Turkish than in ordinary practice. The Swedish language also contains a high frequency of compound forms with various augments between the words. In some cases, these augments may be omitted or they may be the letter «S», «E», «A», «U» or «O» (e.g., "flickebarn" built as "flick+E+barn" (female child)). Also found in this language are more homographs (words having different meanings or interpretations) than in English (e.g., bank) and part-of-speech tagging may be useful in such circumstances in order to remove some keyword ambiguities (Hedlund *et al.*, 2001).

In order to analyze the impact of compound splitting in IR, Kraaij & Pohlmann (1996) evaluated various linguistic and non-linguistic stemming procedures for Dutch, a Germanic language. They found that the use of a compound splitting strategy based on a readable Dutch dictionary seems to improve the basic inflectional stemmer, although the enhancement was not significant. Moreover, they also noted that 40% of the unique word forms were not included in the dictionary (mainly proper nouns and nominal compounds, and to a lesser extent spelling mistakes).

All these various languages thus require different stemming procedures, and our approach is to suggest simple, all-purpose stemmers that can identify noun or adjective variants and mainly ignore variations in verb form. It is our opinion that we should be mostly concerned with nouns and adjectives when building document or query representatives, assigning secondary importance to verbs and their more complicated morphological variations. Moreover, the derivational suffixes (e.g., «-ably», «-ship» present in the English language) were not removed in other languages because when evaluating French document collections it did not result in any clear and significant improvement in retrieval effectiveness (Savoy, 1999).

2.3. Indexing and searching strategies

In order to define a retrieval model, we will first explain how documents and queries are represented and then how these representations are compared, thus resulting in a ranked list of retrieved items. Moreover in this section and in the following, we will describe and evaluate various search strategies in order to present a broader overview of the relative merit of different search models on the one hand, and on the other to ground our conclusions on more firm evidences. Our experiments are carried out with the variant term weighting

schemes of the vector space model and one variant of the probabilistic model, i.e. the Okapi model.

As a first approach, we adopted a binary indexing scheme within which each document or request are represented by a set of keywords without any weight. To measure the similarity between documents and requests, we count the number of common terms, computed according to the inner product (retrieval model denoted "doc=bnn, query=bnn"). See Salton & Buckley (1988) or Appendix 1 for details.

Binary logical restrictions are often too limiting for document and query indexing. It is not always clear whether or not a document should be indexed by any given term, meaning a simple "yes" nor "no" is insufficient. In order to create something in between, the use of term weighting allows for better term distinction and increases indexing flexibility. As noted previously, the similarity between a document and the request is based on the number of terms they have in common, weighted by the component tf (retrieval model notation: "doc=nnn, query=nnn").

In a third IR model (Salton and Buckley, 1988), those terms that do occur very frequently in the collection are not believed to be too helpful in discriminating between relevant and non-relevant items. Thus we might count their frequency in the collection, or more precisely the inverse document frequency (denoted by idf), resulting in a larger weight for sparse words and a smaller weight for more frequent ones. In this case, higher weights are given to terms appearing more often in a document (tf component) and rarely in other articles (idf component). As such, each term does not have an equivalent discrimination power, and a match on a less widely used keyword must therefore be treated as being more valuable than a match on a more common word. Moreover, using a cosine normalization (retrieval model notation: "doc=ntc, query=ntc") may prove beneficial and each indexing weight may vary within the range of 0 to 1.

Other variants may also be created, especially given that the occurrence of a particular term in a document is a rare event. Thus, it may be a good practice to assign more importance to the first occurrence of this word as compared to any successive, repeating occurrences. Therefore, the tf component may be computed as the ln(tf) + 1.0 (retrieval model notation: "doc=ltc, query=ltc") or as $0.5 + 0.5 \cdot$ [tf / max tf in a document]. In this latter case, the normalization procedure is obtained by dividing tf by the maximum tf value for any term in the document (retrieval model denoted "doc=atn"). Different weighting

formulae may of course be used for documents and requests, leading to other different weighting combinations.

Finally we should consider that a term's presence in a shorter document provides stronger evidence than it does in a longer document. To account for this, we integrate document length within the weighting formula, leading to more complex IR models; for example, the IR model denoted by "doc=Lnu" (Buckley *et al.*, 1996), "doc=dtu" (Singhal *et al.*, 1999). In these schemes a match on a small document will be treated as more valuable than a match on a longer document.

If in the vector space model, documents and queries are represented by vectors, in the probabilistic model (Robertson and Sparck Jones, 1976; van Rijsbergen, 1979, Chapter 6), documents and requests representation together with the decision to retrieve documents will be based on the probabilistic theory. Within this framework, various probabilistic models have been suggested, and in this paper, we will evaluate a particular model denoted "Cornell version of BM25", part of the Okapi family (Robertson *et al.*, 2000). In order to simplify our notation, we will refer to this model as "Okapi" (see Appendix 1 for details).

Given that French and Italian morphology is comparable to that of English, we decided to index French and Italian documents based on word stems. For the German language and its more complex compounding morphology, we decided to use a 5-gram approach (see Section 2.5). The question then arises is: "How will these retrieval models behave when used with our corpora?"

2.4. Evaluation of various monolingual corpora

As a retrieval effectiveness indicator, we adopted non-interpolated average precision as a retrieval effectiveness measure (computed on the basis of 1,000 retrieved items per request by the TREC-EVAL program), allowing for both precision and recall to use a single number (Voorhees and Harman, 2000). A decision rule is required to determine whether or not a given search strategy is better than another. The following rule of thumb could serve this purpose: a difference of at least 5% in average precision is generally considered significant and a 10% difference is considered material (Sparck Jones and Bates, 1977, p. A25).

For a more precise decision methodology, we might also apply statistical inference methods such as Wilcoxon's signed rank test or Sign test (Salton and McGill, 1983, Section 5.2; Hull, 1993) or hypothesis testing based on bootstrap methodology (Savoy, 1997). In this

paper, we will base our statistical validation on the bootstrap approach because this methodology does not require that the underlying distribution of the observed data follows the normal distribution. As stated in (Salton and McGill, 1983) and demonstrated in (Savoy, 1997), this hypothesis is not respected, and thus it may invalidate the underlying statistical test. Moreover, the bootstrap methodology may allow us to derive approximate confidence intervals if needed.

In statistical testing, the null hypothesis H_0 states that both retrieval schemes produce similar performance. Such a null hypothesis plays the role of a devil's advocate, and this assumption will be accepted if two retrieval schemes return statistically similar means, and rejected if not. Thus, in the tables found in this paper we have underlined statistically significant differences based on a one-sided non-parametric bootstrap test, based on those means having a significance level fixed at 5%. However, a decision to accept H_0 is not equivalent to the opinion that the null hypothesis H_0 is true, but instead represents the fact that "H_0 has not been shown to be false" resulting in insufficient evidence against H_0.

	Average precision (% change)			
Query (Title only) Model	English 33 queries 579 relevant	French 34 queries 528 relevant	Italian 34 queries 338 relevant	German 37 queries 821 relevant
doc=Okapi, que=npn	37.26	41.62	33.98	31.42
doc=Lnu, query=ltc	32.69 (-12.3%)	36.59 (-12.1%)	32.47 (-4.4%)	27.66 (-12.0%)
doc=atn, query=ntc	31.40 (-15.7%)	39.04 (-6.2%)	28.96 (-14.8%)	31.30 (-0.4%)
doc=dtu, query=dtc	31.96 (-14.2%)	37.89 (-9.0%)	31.04 (-8.7%)	28.23 (-9.8%)
doc=ltn, query=ntc	25.28 (-32.1%)	36.56 (-12.2%)	31.90 (-6.1%)	28.22 (-10.2%)
doc=ntc, query=ntc	18.11 (-51.4%)	25.02 (-39.9%)	20.35 (-40.1%)	23.42 (-25.5%)
doc=ltc, query=ltc	16.76 (-55.0%)	25.09 (-39.7%)	18.39 (-45.9%)	21.51 (-31.5%)
doc=lnc, query=ltc	17.70 (-52.5%)	23.19 (-44.3%)	21.25 (-37.5%)	21.65 (-31.1%)
doc=bnn, query=bnn	12.54 (-66.3%)	22.85 (-45.1%)	19.63 (-42.2%)	23.44 (-25.4%)
doc=nnn, query=nnn	9.69 (-74%)	14.56 (-65.0%)	15.15 (-55.4%)	9.78 (-68.9%)

Table 3: Average precision of various indexing and searching strategies based on monolingual requests and documents

The results in Table 3 show that the Okapi probabilistic model provides the best performance, significantly better than the vector-scheme ("doc=Lnu, query=ltc"). The IR models "doc=atn, query=ntc" or "doc=dtu, query=dtc" perform well, yet not as well as the Okapi search approach. However, based on the bootstrap test, the difference cannot always be viewed as significant (significance level of 5%). A closer look at the Table 3 data demonstrates that, for the German collection and comparing the Okapi IR model with the "doc=ltn, query=ntc" vector-processing scheme, the mean difference is 10.2% and favors the

Okapi approach. The bootstrap test however cannot detect a statistically significant difference. A query-by-query analysis reveals that the Okapi probabilistic model improves retrieval effectiveness for 20 queries out of a total of 37, and for these 20 queries, the mean average precision is improved by 11.95%. On the other hand, for 16 requests, the "doc=ltn, query=ntc" search model produces a better retrieval performance (mean improvement of 7.52%), and for one request, the average precision is the same. In order to find a statistically significant difference between two retrieval schemes, the difference between individual request performance must favor one given retrieval model for a large number of queries on the one hand, and on the other, the difference in performance must be significant (e.g., an improvement of 0.1% will not be very useful).

Finally, the traditional tf-idf weighting scheme (doc=ntc, query=ntc) does not exhibit very satisfactory results, and the simple term-frequency weighting scheme ("doc=nnn, query=nnn") or the simple coordinate match ("doc=bnn, query=bnn") results in poor retrieval performance.

	Average precision (% change)		
Query Model / Mean indexing terms	Title 2.88 terms	Title-Desc 7.95 terms	Title-Desc-Narr 16.9 terms
doc=Okapi, que=npn	41.62	46.29 (+11.2%)	46.73 (+12.3%)
doc=Lnu, query=ltc	36.59	40.11 (+9.6%)	42.17 (+15.2%)
doc=atn, query=ntc	39.04	41.68 (+6.8%)	43.98 (+12.7%)
doc=dtu, query=dtc	37.89	40.12 (+5.9%)	43.93 (+15.9%)
doc=ltn, query=ntc	36.56	38.93 (+6.5%)	41.09 (+12.4%)
doc=ntc, query=ntc	25.02	27.40 (+9.5%)	29.65 (+18.5%)
doc=ltc, query=ltc	25.09	29.10 (+16.0%)	30.78 (+22.7%)
doc=lnc, query=ltc	23.19	26.73 (+15.3%)	32.04 (+38.2%)
doc=bnn, query=bnn	22.85	16.55 (-27.6%)	13.76 (-39.8%)
doc=nnn, query=nnn	14.56	14.09 (-3.2%)	13.69 (-6.0%)

Table 4: Average precision of various monolingual search models using different query formulations (French collection, 34 queries)

For longer requests however these findings may be altered. To analyze this proposition, Table 4 demonstrates the impact of query length on search performance improvement, listing three different query formulations: (1) Title section only, (2) both the Title and Descriptive sections or (3) all three sections (Title, Descriptive and Narrative). Table 4 shows that retrieval effectiveness is enhanced when topics include more search terms, leading to significant enhancement, when comparing retrieval schemes and using queries based on the Title section with those built using the Title, Descriptive and Narrative sections. This finding does not hold however when there is a simple coordinate match ("doc=bnn, query=bnn") or a

simple term-frequency weighting scheme ("doc=nnn, query=nnn"), thus tending to demonstrate that search keywords extracted from the Descriptive or Narrative sections are less likely to discriminate. As shown in Appendix 3, similar conclusions can be drawn with more statistically-based evidence when considering Italian, German and English collections.

When comparing two or more retrieval schemes, the use of overall statistics such as average precision may hide performance irregularities among requests. Based on our query sets, Table 5 depicts, for each retrieval scheme and each collection, the number of best individual runs on a per query basis. When multiple retrieval schemes return the same best retrieval performance, we simply count the fraction (1/# of runs) for each IR model. Thus, for 41.609 queries out of 138 (or for 30%), the best choice is the Okapi strategy which in Table 3 was shown to represent the best IR model. It is interesting to note that the three best vector-space approaches ("doc=Lnu, query=ltc", "doc=atn, query=ntc" and "doc=dtu, query=dtc") provide the best results for 9.775 to 22.109 queries out of 138. This data also shows that even a simple retrieval scheme such as the simple binary indexing scheme ("doc=bnn, query=bnn") represents the best scheme for 9.1 out of 138 requests (or 6.6%). Finally, other vector processing schemes ("doc=ntc, query=ntc" or "doc=ltc, query=ltc") do not in general perform very well.

	Best IR scheme for # of queries				
Query (Title only) Model	English 33 queries	French 34 queries	Italian 34 queries	German 37 queries	Total
Okapi, que=npn	15.143	11.2	7	8.266	41.609
doc=Lnu, que=ltc	6.143	1.2	2.166	0.266	9.775
doc=atn, que=ntc	3.143	2.7	6	10.266	22.109
doc=dtu, que=dtc	1.143	5.7	6.166	2.266	15.275
doc=ltn, que=ntc	3	7.2	6.166	7.1	23.466
doc=ntc, que=ntc	0.143	1	0.166	2.6	3.909
doc=ltc, que=ltc	0.143	1.5	0.166	1.266	3.075
doc=lnc, que=ltc	1.143	0.5	1.166	0.766	3.575
doc=bnn, que=bnn	2	2	3	2.1	9.1
doc=nnn, que=nnn	1	1	2	2.1	6.1
Total	33	34	34	37	138

Table 5: Characteristics of individual retrieval schemes (Title only)

2.5. Indexing German documents

Intuitively, for European languages it seems natural to index documents and queries based on words (or noun phrases) because these languages have clear word boundaries and each word (or phrase) may convey a meaning. For the Chinese language however, each sentence does not reveal word boundaries between ideographs and previous studies have shown that the use of n-grams represents an efficient approach for indexing documents and making queries in this language. For Chinese, it seems more appropriate to consider bigrams since most words are composed of two characters. For example, Nie and Ren (1999) indicate that in their Chinese dictionary bigrams represents 63.3% of the entries. Moreover, word segmentation in Chinese is also problematic because for a given sentence, several legitimate segmentations are sometimes possible.

Thus, for the German language, given its high frequency of compound construction, we decided to use a 5-gram approach (Mayfield *et al.*, 2000; McNamee *et al.*, 2001). This value of 5 was chosen for two reasons: 1) it results in better performance, and 2) it is closer to the mean word length in our German corpus (mean word length: 5.87, standard error: 3.7), and from the data in Table 6, we may conclude that our word indexing strategy is significantly less effective than the 5-gram approach. However, when considering longer requests based on the Title, Description and Narrative sections (see Appendix 3), retrieval performance for the 5-gram approach is still higher than the word-based indexing scheme but the differences are not always significant.

If we were to consider 5-gram indexing and word-based document representations to be distinct and independent sources of evidence about document content, it would be a good practice to combine these two indexing schemes. To achieve this, we add the similarity values obtained by each document extracted from the two separate retrieval models. As shown in the last column of Table 6, this simple combination provides the best performance in our context and it has already been suggested as one of the best combined approach when using other test-collections (Fox and Shaw, 1994; Lee, 1997; Savoy *et al.*, 1996). To resolve the same problem, Manmatha *et al.* (2001) demonstrate how score distribution can be modeled using an exponential distribution for the set of non-relevant documents and a normal distribution for the set of relevant items. These authors show how a mixture model (Bishop, 1995, Section 2.6) of these distributions may be used to map the score to probabilities, and thus to combine different search engines results working with the same document collection,

using these estimated probabilities of relevance and non-relevance. As for other combination approaches, the combined result list only provides better performance when combining "good" retrieval schemes.

As shown in Table 6, retrieval performance differences between the combined approach and 5-gram indexing strategy are not always statistically significant. Moreover, combining the two sources of evidence leads to increased query processing time and space requirements for storing the inverted files. For these reasons, in the rest of this paper we will only consider the 5-gram approach for the German language.

	Average precision (% change)		
Query (Title only) Model	German 5-grams	German words	German combined
doc=Okapi, que=npn	31.42	25.44 (-19.0%)	33.48 (+6.6%)
doc=Lnu, query=ltc	27.66	22.65 (-18.1%)	29.07 (+5.1%)
doc=atn, query=ntc	31.30	21.87 (-30.1%)	33.37 (+6.6%)
doc=dtu, query=dtc	28.23	22.54 (-20.2%)	29.89 (+5.9%)
doc=ltn, query=ntc	28.22	21.40 (-24.2%)	29.32 (+3.9%)
doc=ntc, query=ntc	23.42	15.26 (-34.8%)	23.27 (-0.6%)
doc=ltc, query=ltc	21.51	15.75 (-26.8%)	20.80 (-3.3%)
doc=lnc, query=ltc	21.65	14.86 (-31.4%)	20.98 (-3.1%)
doc=bnn, query=bnn	23.44	16.30 (-30.5%)	25.10 (+7.1%)
doc=nnn, query=nnn	9.78	10.44 (+6.7%)	10.35 (+5.8%)
Mean difference		-22.83%	+3.39%

Table 6: Average precision of various indexing and searching strategies based on monolingual requests and documents (German collection, 37 queries)

Finally, based on this experiment, we cannot conclude that 5-gram is the only approach to consider when indexing and retrieving documents in German (or in the Swedish, Norwegian, Danish or Dutch languages, which also contain a large number of compound constructions). Further studies are needed in order to compare the n-gram indexing scheme with the heuristic word segmentation approaches for the German language. In the same vein, Nie and Ren (1999) have shown that appropriate word segmentation in Chinese may result in better retrieval performance, less processing time during retrieval and less storage space than does the simple bigram approach.

Other approaches using a German dictionary have already been suggested (Braschler and Schäuble, 2001), resulting in average precision of 40.30 (queries built with the Title, Descriptive and Narrative sections, an IR model based on "doc=Lnu, query=ltu"). Using a 5-gram approach and the same query construction, we obtained an average precision of 40.17

(search model "doc=Okapi, query=npn", see Appendix 3, Table A.7) and an average precision of 42.09 when combining the 5-gram scheme with word-based indexing strategy.

3. Bilingual information retrieval

In the previous chapter, we obtained a better understanding of language-dependant retrieval approaches, showing that search models that work well in English also perform well in other languages. Since we live in a multilingual world however, automatic information retrieval should be designed with this constraint in mind. We will thus take a look at bilingual search models that, based on queries written in English, can retrieve relevant information from document collections written in French, Italian or German (Oard and Dorr, 1996; Grefenstette, 1998). To cross this language barrier, we have based our approach on free and readily available translation resources that can automatically provide complete translations of queries submitted in the desired target language.

The first section of this chapter describes some related works while Section 3.2 presents our combined strategy and compares the retrieval effectiveness of our approach to other solutions proposed. Section 3.3 describes the pseudo-relevance feedback used to hopefully improve the retrieval performance by reducing the underlying ambiguity of all machine-based translation. Section 3.4 provides a query-by-query analysis of the relative merit of various query translation strategies. Finally, in Section 3.5, we analyze how retrieval effectiveness can vary when comparing various manually-based requests translations from English to French.

3.1. Related work

In an early work, Salton (1971) showed that by using a thesaurus that was carefully constructed manually, cross-language retrieval could be as effective as that of monolingual retrieval. In this case, each thesaurus class groups related words in both German and English, thus serving as an interlingual link between English terms and their related German translations. During indexing, the system searches for the thesaurus class associated with the document's terms and replaces them with the corresponding thesaurus class identifier. In building such a bilingual thesaurus however the matching of English words to their German equivalent (or other languages) is not always perfect. For example, the English term "bank" may be translated in German by "Bank" (in the context of financial institutions) or by "Flussbank" (for river bank). Moreover some words do not appear in the thesaurus, a

problem often encountered with other dictionary look-up procedures, and especially for proper nouns. Finally, collections of document abstracts were very small compared to the current standard (1,095 abstracts in English, 468 in German). Derived from this work, we might consider manually indexing case law using a manually built controlled thesaurus. This solution is used by the Swiss Supreme Court's retrieval system.

Based on parallel corpora, Landauer and Littman (1990) suggested using latent semantic indexing (Furnas *et al.*, 1988) to generate a multidimensional indexing space for English documents and their French translations. The test-collection used in this study was relatively small and no retrieval effectiveness evaluations were conducted. Moreover, the parallel collections (defined as corpora where the same documents are presented in two or more languages) were not always available.

Sheridan and Ballerini (1996) suggested generating co-occurrence information from comparable corpora, but unaligned, in order to find statistically related terms in the target language for a better translation quality. The average precision obtained by this similarity thesaurus is still considerably below that of single-language retrieval. Moreover, comparable corpora were not readily available. To partially resolve this problem, Nie *et al.* (1999) suggested using their PTMiner system to extract parallel corpora from the Web. Then using these Web page collections, sentences from two pages written in two different languages were aligned using a length-based alignment algorithm (Gale and Church, 1993). The system then computed the probabilities of translating one term into another (using an expectation maximization principle (Brown *et al.*, 1993)). With this type of statistical translation model, quality of sources (e.g., Web sites) and the size of available corpora were of prime importance (Nie and Simard, 2001).

Franz *et al.* (1999) suggested a similar approach to translate the documents rather than queries and this approach provided very interesting performance in the cross-language track at TREC-7. Cultural, thematic and time differences may also play a role in the effectiveness of such approaches (Kwok *et al.*, 2001). In the same vein, Xu & Weischedel (2001) demonstrated that a more effective translation of requests might be obtained using a combination of four different bilingual lexicons, in part derived from parallel corpora. However, the parallel collection closest to the test-collection usually provided the best retrieval effectiveness.

Bilingual machine-readable dictionaries may also be considered. In this vein, Hull and Grefenstette (1996) proposed an approach that resulted in a drop of 50 % in average precision when translating words based on such bilingual dictionaries, compared to monolingual performance. These authors found that the correct identification and translation of multi-word expressions can make the biggest difference in average performance, compared to the problem of resolving translation ambiguity (e.g., "bank" translated as "financial institution" or as "river bank") or when faced with missing terminology (e.g., a given word does not appear in the bilingual dictionary).

Pirkola (1998) showed that structured queries may resolve some problems related to the presence of multiple translation alternatives when using bilingual dictionaries to translate the submitted request. Within structured queries (e.g. using the INQUERY system (Broglio *et al.*, 1995)), we could distinguish between search keywords having the same influence (as we have done in this paper) and the presence of synonyms (various words expressing the same topic facet) using a special operator (syn-operator). Moreover, a third operator could be used as a proximity operator. In this system, translation alternatives derived from the same word can be grouped under the same syn-operator, and the query might thus include disjunctive relationships between search keywords. On the other hand, the proximity operator can be applied to represent phrases or compound constructions that are translated using multiple words. The evaluation of this approach as described in Pirkola (1998) shows results close to those achieved by a monolingual system, and the use of structured queries is also described in Hiemstra *et al.* (2001).

In order to limit the translation ambiguity, David and Ogden (1997) suggested combining information extracted from dictionaries with shallow natural language processing, such as part-of-speech tagging and phrase recognition, during the interactive query translation process. Similarly, Oard and Resnik (1999) described experiments in which the user may disambiguate words fairly easily when the system provides up to three alternative translations in a short context. In our approach, we adopted a fully automatic approach that does not require the user to select the more appropriate translation alternative. In the same vein, Kraaij *et al.* (2000) also used a bilingual dictionary to automatically translate the request. They suggested weighting the translation alternatives based on the number of senses that could be found in the dictionary. Finally, they also suggested expanding the request using a lexical database storing synonyms, hyperonyms (generalization as "Ford" and "car") and hyponyms ("bank" and "banker", "deposit" or "loan") relationships between words.

Ballesteros and Croft (1998) suggested using a bilingual dictionary to perform word-by-word translation and then adding terms to the query through pre-translation query modification, using pseudo-relevance feedback and post-translation request expansion. The retrieval effectiveness for English-Spanish cross-language retrieval was attractive but still below that obtained for monolingual collections.

Braschler and Schäuble (2001) suggested using available machine translation software to automatically translate queries, documents or both. In this study, the document translation-based approach performed better than the query translation-based retrieval scheme. Moreover, the combination of similarity thesaurus, document and query translation-based resulted in the best performance. The average precision was still below that obtained by a monolingual search, and when searching on the Web, such an approach may not be the most appropriate answer. Thus query translation seems to be more realistic. Knowing that requests are typically much shorter than documents, it is generally more efficient to translate just the query rather than translate each document, particularly if the search system must retrieve documents written in several languages.

Statistical language modeling (Ponte and Croft, 1998; Xu and Croft, 1999) can also be used in the context of bilingual or cross-language information retrieval. In these IR models, the retrieval process infers the use of a language model for each document and estimates the probability of generating the submitted request according to each of these models. The documents are then ranked according to these probabilities. Such an approach was suggested by Kraaij (2001) and Hiemstra *et al.* (2001) within which the translation probabilities of a given term into the target language is estimated, based on parallel corpora.

Finally, we must mention that a bilingual IR system approach could be based on a hidden Markov model (Miller *et al.*, 1999) as suggested by Xu & Weischedel (2001). In this case, the underlying transition probabilities in their Markov model were estimated, based on a bilingual dictionary (uniform distribution) and on parallel corpora. Using pseudo-relevance feedback both before and after translating the request (the first query expansion was performed in the original language and the second on the target language), these authors showed that the resulting bilingual performance outperforms the retrieval effectiveness achieved by the monolingual run.

3.2. Bilingual information retrieval

In our bilingual experiments, we were faced with the following situation. We used the English set of queries provided in the CLEF 2000 test suite (see Appendix 2, Table A.2) but we did not have any parallel or aligned corpora from which to derive statistically or semantically related words in the target language. In order to develop a fully automated approach, we chose to translate the requests using the SYSTRAN™ system (Gachot *et al.*, 1988, available for free at http://www.systran.com) and also to translate query terms word-by-word using the BABYLON bilingual dictionary (available at http://www.babylon.com). The BABYLON bilingual dictionary might suggest not only one, but several candidates for each word, thus revealing the underlying ambiguity of a given term.

Of course, various errors can result from automatically translating a query formulation, due to the bilingual dictionary's limited coverage, the term's underlying ambiguity, the correct identification of multi-word concepts and their appropriate translation, and the translation of proper names. For example, for the word "Electroweak" or for the term "privatisation" were not found in the BABYLON dictionary. These terms were left untranslated in the translated queries. Further examples of queries along with their successful and unsuccessful translations are given in Appendix 2.

When translating the query "solar temple" into German, the BABYLON dictionary suggests "Heiligtum" (in the sense of a holy place, which is not absolutely wrong) but the appropriate translation is "Sonnentempel". For this request, the SYSTRAN system proposed "SolarBügel" (solar captor) which is clearly wrong. The underlying ambiguity of terms generates other errors. For the request "sinking of the Estonia", the SYSTRAN system selects the wrong French translation of the word "sinking" while the BABYLON bilingual dictionary gives the correct French term.

In order to obtain a quantitative picture of a term's ambiguity, we analyzed the number of translation alternatives generated by BABYLON'S bilingual dictionaries. For this study, we do not consider determinants (e.g., "the"), conjunctions and prepositions (e.g., "and", "in", "of") or words appearing in our English stopword list (e.g., "new", "use"), terms that generally having a larger number of translations. Based on the Title section of the English requests, we found 106 search keywords to be translated.

From the data depicted in Table 7, we can see that the mean number of translations provided by BABYLON dictionaries varies according to language, from 2.96 for German to

5.88 for Italian. We found the maximum number of translation alternatives for the word "wind" in French and German and for the term "single" in Italian. The median values of these distributions is rather small, varying from 1.6 for German to 3.4 for Italian. Thus when considering the first two translation alternatives, we covered more than 50% of the keywords to be translated in German, 50% in French and 39.6% for the Italian language. Figure 1 provides a clearer picture of how the number of translation alternatives is relatively concentrated around one.

	Number of translation alternatives		
Query (Title only)	French	Italian	German
Mean number of translations	3.60	5.88	2.96
Standard deviation	3.60	5.73	2.84
Median	2.1	3.4	1.6
Maximum	20	26	18
No translation	3	3	3
Only one alternative	32	30	36
Two alternatives	17	9	23
Three alternatives	17	9	17

Table 7: Number of translations given by the Babylon system for the English keywords appearing in the Title section of our queries

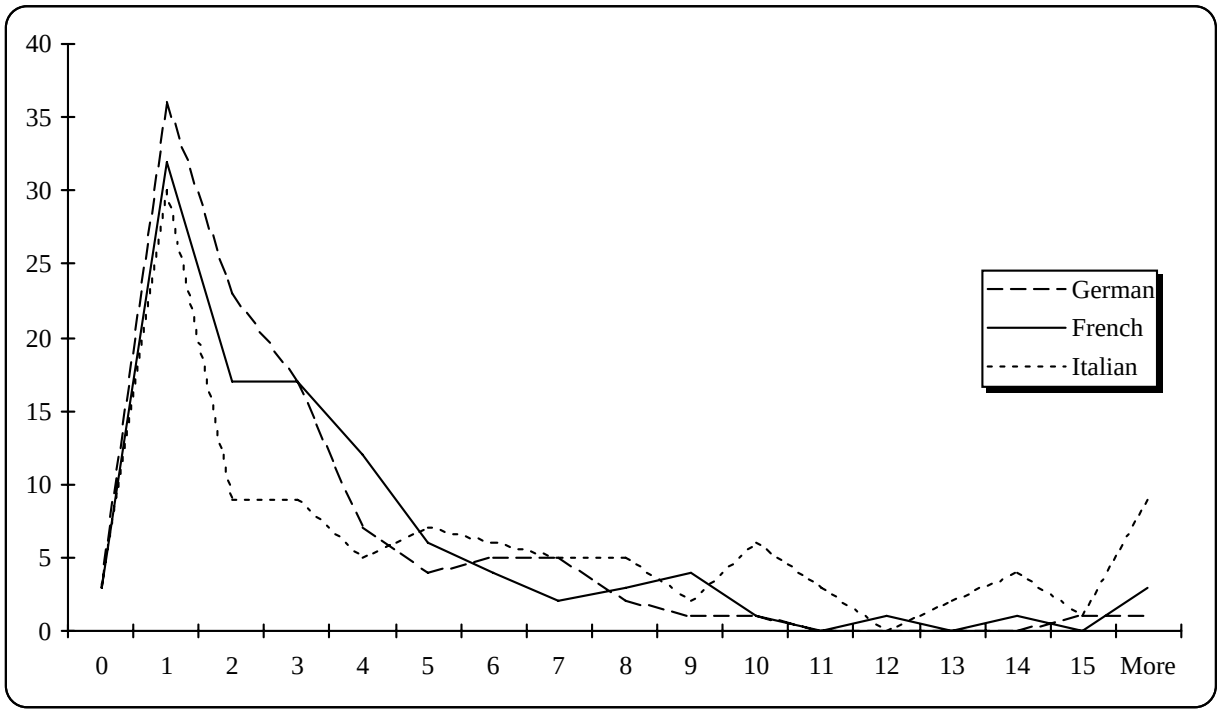

Figure 1: Distribution of the number of translation alternatives

Failing to recognize multi-word concepts is also another source of performance degradation because the automatic translation approaches do not produce the best translation.

In Italian for example, the automatic translations of the request "use of wind power" (see Appendix 2, Table A.3) does not lead to the retrieval of many relevant documents. On the other hand, the manual translation of the expression ("energia eolica") extracts many relevant articles from the collections.

In the CLEF 2000 query set, we did not encounter numerous difficulties with proper names because they were very common (e.g., "European", "French", "Holland") or they appeared in the same form across languages (e.g., "Berlin", "Pierre Bérégovoy", "Nobel"). Of course, not all automatic translations were correct. For example, the term "US" must be translated as "USA" in German, but the BABYLON bilingual dictionary produced "uns" (the term "US" was analyzed as a pronoun) and SYSTRAN system gave "us" (no modification). This example demonstrates that both translation tools are generally not case sensitive. Finally, in some circumstances, it is not appropriate to translate a given proper noun; for example, the request "sinking of the Estonia", "Estonia" corresponds to the boot name and translating this noun as "Estonie" (in French) decreases retrieval performance (in the relevant articles, the name "Estonia" appears but not the translated form). When looking up the city name "Nice", the bilingual dictionary treats this proper name as an adjective.

Given that we know that automatic translation tools do not always produce the most appropriate translations, do these errors result in poor retrieval performance? To answer this question, we decided to automatically translate the English requests using the SYSTRAN machine translation system (retrieval performance shown in Table 8 under the heading "SYSTRAN"), and as an alternative, to evaluate the performance achieved by the BABYLON bilingual dictionary. Since the latter suggests different translation alternates, we picked the first translation available (performance depicted under "BABYLON 1") or the first two terms (listed under heading "BABYLON 2"). Our choice was based on the assumption that the more appropriate translation will appear in the first (or in the second) position in the response list provided by the bilingual dictionary. Moreover, our previous analysis of the number of translation alternatives indicated that when considering both translations, we usually cover most of the dictionary entries. However, other studies suggested that including all translation terms (Xu & Weischedel, 2001), or selecting either the most appropriate or a limited number of translation alternatives (e.g., the alternative appearing most frequently in the test collection (Chen *et al.* 2001), or a maximum of six candidates (Kwok *et al.*, 2001)); under the assumption that the greater a term's occurrence in a collection, the greater the probability that it is a good translation.

As reported in previous work, our experiments shown in Table 8 indicate that the average precision produced by using machine-translation systems or a bilingual dictionaries is clearly below that achieved by monolingual runs (average precision listed under the heading "Monolingual"). Using two translated words instead of one decreases the retrieval effectiveness to a greater extent (-47.65% in average for the French collection). When we compared Italian and German bilingual retrieval performances (see Appendix 4), we were able to draw similar conclusions. Of course, the absolute performance values were not the same when considering the other two languages, but the retrieval effectiveness obtained with the SYSTRAN system approach was clearly better than that from the BABYLON bilingual dictionary.

	Average precision (% change)				
Model	Monolingual	SYSTRAN	BABYLON 1	BABYLON 2	Combined
Okapi-npn	41.62	29.64 (-28.8%)	29.45 (-29.4%)	21.37 (-48.7%)	33.14 (-20.4%)
Lnu-ltc	36.59	25.64 (-29.9%)	24.44 (-33.2%)	20.40 (-44.2%)	28.34 (-22.6%)
atn-ntc	39.04	26.37 (-32.5%)	28.58 (-26.8%)	20.15 (-48.4%)	30.87 (-20.9%)
dtu-dtc	37.89	26.66 (-29.6%)	28.99 (-23.5%)	21.13 (-44.2%)	30.01 (-20.8%)
ltn-ntc	36.56	25.23 (-31.0%)	26.79 (-26.7%)	21.06 (-42.4%)	29.80 (-18.5%)
ntc-ntc	25.02	14.69 (-41.3%)	14.93 (-40.3%)	13.08 (-47.7%)	17.14 (-31.5%)
ltc-ltc	25.09	14.42 (-42.5%)	13.83 (-44.9%)	11.78 (-53.1%)	16.41 (-34.6%)
lnc-ltc	23.19	13.90 (-40.1%)	12.66 (-45.4%)	11.56 (-50.2%)	16.14 (-30.4%)
bnn-bnn	22.85	14.59 (-36.1%)	12.10 (-47.1%)	8.49 (-62.8%)	16.69 (-30.0%)
nnn-nnn	14.56	8.60 (-40.9%)	10.14 (-30.4%)	9.49 (-34.8%)	10.88 (-25.3%)
Mean difference		-35.28%	-34.75%	-47.65%	-25.19%

Table 8: Average precision for various translating strategies using the French collection (Title only, using 34 English queries)

In order to improve search performance, we tried combining the machine translation approach results with those produced by a bilingual dictionary. In order to verify this hypothesis, we added the first translated word produced by a bilingual dictionary look-up to each translated query generated by the SYSTRAN system. The average precision obtained from this combined strategy is shown under the heading "Combined" in Table 8, illustrating how this approach performs better than the other bilingual search strategies. However, the average precision was still below that obtained by manually translating the requests into French. For the Italian and German corpora, we could be able to draw identical conclusions (see Appendix 4).

As shown in Table 8, the most effective approach to translating request appears to be a combination of various translation schemes. In this vein, we might translate queries using different machine translation systems (Jones & Lam-Adesina, 2001), add translation

alternatives given by one (or more) bilingual dictionaries (Kwok *et al.*, 2001), using machine translation improved through using terms extracted from wordlists built automatically based on parallel corpora (Kraaij, 2001), bilingual dictionaries and statistical translation models (Nie & Simard, 2001) or extracting manually related terms taken from various Web pages (e.g., using a search engine and inspecting the top ten retrieved documents) (Chen *et al.*, 2001). This additive approach might be combined with a pruning procedure that attempts to remove useless alternatives or reduces search keyword ambiguity (McNamee and Mayfield, 2001; Chen, 2001; Nie & Simard, 2001). However, all these query translation technique combinations may not perform better than any individual tool. For example, Adriani (2001) reports that combining a bilingual dictionary with a parallel corpus does not perform better than using only a bilingual dictionary. Similar findings are reported by Kwok *et al.* (2001) and Xu & Weischedel (2001).

3.3. Pseudo-relevance feedback

It has been observed that pseudo-relevance feedback (blind expansion) is a useful technique for enhancing retrieval effectiveness. For example, we evaluated the Okapi search model with and without query expansions in order to verify whether or not this technique might improve retrieval performance when using various query formulations. In this study, we adopted Rocchio's approach (Buckley *et al.*, 1996) with $\alpha = 0.75$, $\beta = 0.75$ where the system was allowed to add to the submitted query 10 search keywords, extracted from the 5-best ranked documents. The resulting retrieval effectiveness is depicted in the top half of Table 9 for monolingual collections and in the bottom half for bilingual retrieval. We also tuned the parameters of this blind query expansion, as illustrated in the "best performance" row, showing the best average precision that could be achieved using this strategy (the corresponding parameter setting is given in the following row).

	Average precision (% change)			
Model / Query	English	French	Italian	German
Okapi-npn	37.26	41.62	33.98	31.42
+ query expansion	35.33 (-5.2%)	41.98 (+0.9%)	40.28 (+18.5%)	34.05 (+8.4%)
Best performance		42.22 (+1.4%)	40.91 (+20.4%)	37.71 (+20.0%)
# doc / # terms		10 doc / 10 t	5 doc / 20 t	5 doc / 75 t
Okapi-npn		33.14	25.78	25.43
+ query expansion		35.83 (+8.1%)	30.76 (+19.3%)	26.69 (+5.0%)
Best performance		36.00 (+8.6%)	32.10 (+24.5%)	28.33 (+11.4%)
# doc / # terms		5 doc / 15 t	5 doc / 30 t	10 doc / 125 t

Table 9: Average precision with blind query expansion

For the Italian and German corpora, pseudo-relevance feedback results are satisfactory and reveal statistically significant enhancement compared to baseline performance. With the English collection however a decrease in average precision exists, although it is not statistically significant, while for the French corpus the improvement was only statistically significant for our bilingual experiments.

In a bilingual context, pseudo-relevance feedback can be used to find more related terms, extracted directly from the target language as suggested by (Ballesteros and Croft, 1998). For our three bilingual retrievals, this technique improved average precision, leading to mean performances of around 35% for the French collection, 32% for the Italian or 27% for the German corpus. More precisely and for the French corpus in particular, blind query expansion (adding 10 terms extracted from the 5-best ranked documents) improves the average precision for 18 requests and decreases the performance for 12 other queries (4 requests depicting identical performance). For the Italian collection, pseudo-relevance feedback (adding 10 terms from 5 documents) increases the performance for 25 queries and degrades the average precision for 8 requests. Comparing these results with performances shown in Table 3 (monolingual retrieval), the resulting differences were small but still below the best monolingual runs.

3.4. Query by query analysis

As we known, overall statistics such as average precision may hide performance irregularities among requests. Thus, in order to obtain a better picture of the relative merit of each query translation-based strategy, we analyzed the average precision of each query obtained with the Okapi probabilistic model (without pseudo-relevance feedback query expansion) and grouped them in three different classes. In the first, we grouped those queries having average precision identical to that of monolingual retrieval (requests provided by the CLEF 2000 test suite). For example, when we submitted the request "architecture à Berlin" (monolingual) or "architecture dans Berlin" (translation provided by SYSTRAN), the same average performance was obtained for both request formulations. In the second group, we grouped queries that improved the average performance when compared to the monolingual run, and for the third group the automatically translated request resulted in a lower retrieval performance. The results obtained using our three corpora and with our three query-translation approaches are listed in Table 10.

Language / System	SYSTRAN	BABYLON 1	Combined
French (34 queries)	16 / 4 / 14	11 / 3 / 20	11 / 7 / 16
German (37 queries)	14 / 7 / 16	4 / 5 / 28	6 / 9 / 22
Italian (34 queries)	8 / 4 / 22	6 / 4 / 24	0 / 9 / 25
Mean difference in average precision for the French collection			
When better	+ 22.44%	+ 24.40%	+ 11.33%
When worse	- 35.53%	- 24.36%	- 22.98%

Table 10: Query-by-query analysis of various translation strategies

From the data shown in the top half of Table 10, for the French collection it is evident that the SYSTRAN system seems more effective than BABYLON approach (we obtained the same average precision for 16 queries, 4 requests providing better average precision when translated by the SYSTRAN system, and for 14 queries, the formulation provided in the CLEF 2000 test collection resulted in better retrieval effectiveness). For the German corpus, 14 queries showed identical performance while 7 requests improved average precision over the monolingual retrieval, and for 16 requests, the requests provided in the CLEF 2000 test suite performed better. For the Italian collection however, 22 requests showed better average precision when using the formulation given by the CLEF 2000 test collection, compared to the query translation produced by the SYSTRAN system.

3.5. Variability in query formulation

So far we have implicitly admitted that manually translating requests will always produce the same query formulation or, at least, a very close formulation that has no real impact on retrieval performance. To investigate this hypothesis, we asked eleven subjects to manually translate into French the English set of queries provided in the CLEF 2000 test suite (Title only). To assist in this process, we provided each of the eleven persons with the requests, including their Description and Narrative logical sections. Given that the titles were relatively short (a mean of 2.6 indexing terms in English) and the context in which the descriptive and narrative parts were relatively unambiguous, it was our opinion that the resulting query translations into French would return very similar queries.

In order to verify this assumption, we evaluated the eleven sets of manually translated queries using ten retrieval schemes. The results are depicted in Table 11 where the second column lists the average precision obtained by the French queries included in the test collection (denoted "CLEF 2000 formulation"). The following columns listed the average precision achieved, in mean, by our eleven query formulations, together with the best and the

worse retrieval performances. From this data, it is clear that the CLEF 2000 formulation resulted in the best retrieval performances (in fact, it was the best formulation for seven retrieval models from a total of 10).

Query formulation has a big impact on retrieval performance and the mean difference between the maximum and the minimum average precision as shown in Table 11 is 10.26. Looking at the standard deviation for our eleven query expressions (last column of Table 11), a relatively large variations across average performance can also be noted. As a rule of thumb, we could say that the average precision achieved by the Okapi probabilistic model was 35.16 ± 1.812 · 3.67 = [41.80 - 28.51] (approximate confidence interval having a coverage probability equal to 90%, built using the Student distribution with 10 degrees of freedom). These findings contradict our assumption, and we must admit that retrieval performance varies a great deal between users, who had to manually translate a given request.

	Average precision				
Model	CLEF 2000 formulation	Average translation	Maximum	Minimum	Standard deviation
Okapi-npn	41.62	35.16 (-15.5%)	42.35	29.29	3.67
Lnu-ltc	36.59	29.89 (-18.3%)	35.35	25.28	2.29
atn-ntc	39.04	31.96 (-18.1%)	39.58	24.68	3.86
dtu-dtc	37.89	32.80 (-13.4%)	39.45	27.36	3.48
ltn-ntc	36.56	30.56 (-16.4%)	36.92	25.09	3.35
ntc-ntc	25.02	18.93 (-24.3%)	24.16	15.48	2.61
ltc-ltc	25.09	18.36 (-26.8%)	23.28	14.98	2.38
lnc-ltc	23.19	17.26 (-25.6%)	21.38	13.59	2.14
bnn-bnn	22.85	16.62 (-27.3%)	21.51	12.52	2.88
nnn-nnn	14.56	10.62 (-27.1%)	14.38	7.50	2.15

Table 11: Average precision of various manually translated query formulations (French collection, 34 queries, Title only)

Based on a query-by-query analysis of the "doc=Okapi, query=npn" search model performance, there are four queries (over a total of 34) whose average precision is identical across all query formulations. These requests do not exhibit real translation difficulties ("The suicide of Pierre Bérégovoy", "World Trade Organization", "Corruption in Italy" or "Cancer genetics"). For seven of the requests, only one subject wrote a query expression that resulted in lower average precision (e.g., the request "architecture in Berlin" was translated by ten persons as "architecture à Berlin" and once as "le style architectural de Berlin", for which the retrieval effectiveness was lower).

On the other hand, we found a set of five queries for which the CLEF 2000 formulation outperformed our eleven query formulations. For example, the request "postmenopausal pregnancy" was translated officially as "grossesses post-ménopausiques". Our subjects provided various phrases such as "grossesse post-ménopausale", or "grossesse après la ménopause" resulting in lower retrieval performances. For this request, the term "post-ménopausiques" managed to retrieve the greatest number of relevant documents, and our simple stemmer was not able to conflate the term "post-ménopausiques" and "post-ménopausale" to the same root. Within this same group was the request "use of wind power" (written in the CLEF 2000 formulation as "utilisation de la force éolienne") which was usually translated word-by-word by our subjects (e.g., "utilisation de la puissance du vent" or "utiliser la puissance du vent"). In an extreme case, one of our subjects submitted the translation "l'aéronautique (ou aérodynamique)" that did not retrieve any relevant documents. However, another person submitted "l'utilisation de l'énergie éolienne" which for this request obtained the best performance.

Overall, in comparing average precision for the 34 requests by our eleven subjects (374 observations) with the CLEF 2000 formulation, we found that retrieval performance was identical in 46.52% (174 observations cases), while the CLEF 2000 formulation was better in 35.83% (134 observations) and worse in 17.65% (66 observations) of the query-by-query evaluations.

Based on these findings, it is our opinion that it would be more realistic to compare retrieval performances from our fully automatic combined query translation approach (Section 3.2, Table 8) with the average performances achieved by our eleven subjects. In Table 12, we reported this mean average precision and the retrieval performance of our combined bilingual retrieval scheme (French collection). The retrieval effectiveness for the human-based translation was still higher but the performance differences were relatively small and the statistical test could not detect any significant difference. However, this experiment was based on a short query formulation (Title only), and if longer request expressions are considered, this picture may change.

Model	Average precision	
	Average translation	Combined
doc=Okapi, query=npn	35.16	33.14 (-5.7%)
doc=Lnu, query=ltc	29.89	28.34 (-5.2%)
doc=atn, query=ntc	31.96	30.87 (-3.4%)
doc=dtu, query=dtc	32.80	30.01 (-8.5%)
doc=ltn, query=ntc	30.56	29.80 (-2.5%)
doc=ntc, query=ntc	18.93	17.14 (-9.5%)
doc=ltc, query=ltc	18.36	16.41 (-10.6%)
doc=lnc, query=ltc	17.26	16.14 (-6.5%)
doc=bnn, query=bnn	16.62	16.69 (+0.4%)
doc=nnn, query=nnn	10.62	10.88 (+2.5%)
Mean difference		-4.90%

Table 12: Average precision of manual vs. automatic query translation approaches (French collection, 34 queries, Title only)

4. Multi-lingual information retrieval

In the previous chapter, we obtained a better understanding of the retrieval effectiveness of various bilingual retrieval approaches. However, this represents only the first step in analyzing cross-language information retrieval systems. In this chapter, we will investigate the situation where users write a request in English in order to retrieve relevant documents in English, French, German and Italian. To deal with this multi-language barrier we have based our approach on solutions described in the previous chapter. Therefore, the different collections were indexed separately using a language specific procedure. In this chapter, we will investigate various database merging strategies that will allow us to present users with a single list of retrieved articles.

The first section of this chapter describes some related works suggesting different result list merging strategies. Section 4.2 will analyze and evaluate various database merging approaches within which queries are manually translated while Section 4.3 will present the relative retrieval performance of these various merging strategies when dealing with automatic machine-based translation schemes.

4.1. Related work

Recent work has suggested various solutions to merge separate results list obtained from separate collections or distributed information services. As a first approach, we will assume that each collection contains approximately the same number of pertinent items and

that the distribution of the relevant documents is similar across the result lists. Based solely on the rank of the retrieved records, we can interleave the results in a round-robin fashion. According to previous studies (Voorhees *et al.*, 1995a; Callan *et al.*, 1995), the retrieval effectiveness of such interleaving scheme is around 40% below that achieved from a single retrieval scheme, working with a single huge collection that represents the entire set of documents. However, this decrease may diminish (around -20%) when using other collections (Savoy and Rasolofo, 2001).

Voorhees *et al.* (1995b; 1996) demonstrated that we may improve this ranking scheme by comparing the estimated expected relevance of each collection to the current request. Thus, instead of extracting an equal amount of records from each collection, the suggested scheme retrieves, for each result list, a number of documents related to the previous performance of the underlying collection. Depending on the underlying learning schemes, the overall performance was 20% to 30% below the average precision produced by a single huge collection. Moreover, approaches based on a training set cannot be usually used with new collection.

To take account for the document score computed for each retrieved item (or the similarity value between the retrieved record and the request), we might formulate the hypothesis that each collection is searched by the same or a very similar search engine and that the similarity values are therefore directly comparable (Kwok *et al.*, 1995; Moffat and Zobel, 1995). Such a strategy, called raw-score merging, produces a final list sorted by the document score computed by each collection. However, as demonstrated by Dumais (1994), collection-dependent statistics in document or query weights may vary widely among collections, and therefore this phenomenon may invalidate the raw-score merging hypothesis.

To account for this fact, we might normalize document scores within each collection by dividing them by the maximum score (i.e. the document score of the retrieved record in the first position). As a variant of this normalized score merging scheme, Powell *et al.* (2000) suggest normalizing the document score rsv_j according to the following formula:

$$rsv'_j = \left(rsv_j - rsv_{min}\right) / \left(rsv_{max} - rsv_{min}\right) \quad (1)$$

in which rsv_j is the original retrieval status value (or document score), and rsv_{max} and rsv_{min} are the maximum and minimum document score values that a collection could achieve for the current request.

As a fourth merging strategy, Callan *et al*. (1995) suggest a merging strategy based on the score achieved by both collection and document. The collection score, denoted s_i for the ith collection, is computed according to the probability that the corresponding collection respond appropriately to the current request. For a given query Q and for each collection, the CORI system computes a collection score defined as:

$$s_i = \frac{1}{m} \cdot \sum_{k=1}^{m} s(t_k \mid C_i)$$

within which m indicates the number of query terms in the current request Q, and $s(t_k \mid C_i)$ indicates the contribution of the search term t_k in the score of collection C_i calculated as follows:

$$s(t_k \mid C_i) = defB + (1 - defB) \cdot \frac{df_i}{df_i + K} \cdot \frac{\ln\left(\frac{|C|+0.5}{cf_k}\right)}{\ln(|C|+1.0)} \quad \text{with } K = k \cdot \left[(1-b) + b \cdot \frac{lc_i}{avlc}\right]$$

in which |C| is the number of collections (four in our case), df_i indicates the number of documents in collection C_i containing the kth query term, cf_k is the number of collections containing the query term t_k, lc_i is the number of indexing terms in C_i, avlc is the average number of indexing terms in each collection, and defB, b and k are constants assigned the following values: defB = 0.4, k = 200 and b = 0.75, as suggested by (Callan *et al*., 1995).

Based on this collection score denoted s_i, the collection weight w_i assigned to the ith collection is:

$$w_i = 1 + |C| \cdot \left[(s_i - \bar{s})/\bar{s}\right] \qquad (2)$$

within which $\bar{s}$ is the mean of collection scores, and |C| is the number of collections as defined above.

The resulting weight w_i will be used to modify the similarity value attached to each document. Instead of directly using this document score (as used in the raw-score merging strategy), the final document score of the jth document belonging to the ith collection is the product of the collection weight w_i by its original document score (that is $rsv'_j = rsv_j \cdot w_i$), and then the CORI system merges the result lists according to these new document scores.

As a fifth strategy, we may use the logistic regression (Flury, 1997, Chapter 7; Hosmer and Lemeshow, 1989) to predict the probability of a binary outcome variable according to a set of explanatory variables. Based on this statistical approach, Le Calvé and Savoy (2000) described how to predict the probability of relevance of documents retrieved by different

retrieval schemes or collections. Instead of the original document score rsv_i, the resulting estimated probabilities will be used in sorting the retrieved records obtained from separate collections, in order to obtain a single ranked list. However, to estimate the underlying parameters, such an approach requires a training set, that might not always be available.

As a sixth approach to merging different result lists, we could account for the fact that these result lists are obtained from different languages. In this vein, Franz *et al.* (1999) and Franz *et al.* (2000) indicated that document scores, even when computed according to the same search engine, are not directly comparable when extracting documents from different collections. Different languages and different qualities of the underlying translation resources do not produce comparable document scores. Thus, these authors suggest estimating the probability that a given document is relevant based on its rank (r), the document language l_d and also upon features of the query Q. Thus for a given document, the probability of relevance is a function of r, l_d, and Q, and in order to estimate this probability of relevance, they observed that for a given rank r the precision is approximately a linear function of the ln(r), as confirmed by Le Calvé & Savoy (2000). Therefore, they suggested estimating this probability as a simple function only of its rank, and estimating this function for each document language l_d. A similar approach was used by Kraaij *et al.* (2000).

Moreover, we may also obtain a separate estimation in accordance with some query features (Franz *et al.*, 2000). As described in Section 2.1, we knew that a significant fraction of the CLEF 2000 queries concerns local events (e.g., for the query "The French Academy" or "Wolves in Italy") that are under-reported in some sources (e.g., European questions marginally reported in the *Los Angeles Times*). For example, the request "Corruption in Italy" owns 26 relevant articles in the Italian documents, 8 in the French corpus and 4 in the German source, and, in the other hand, the request "Teaching techniques for non-English speakers" obtains 25 relevant documents, all from the *Los Angels Times*. To recognize such requests, Franz *et al.* (2000) suggest considering whether or not the submitted query mentions the name of an European country. Upon an inspection of the CLEF 2000 queries (see Appendix 2), we count nine requests mentioning an European country (as a noun or an adjective) or a given city (e.g., "Berlin" or "Nice"). Based on the TREC-8 test collections, taking account of such query features may marginally improve average precision (Franz *et al.*, 2000). When inspecting the distribution of the relevant items in the CLEF 2000 test suite, there are queries for which this prior distinction was not really helpful. For example, the request "Olive oil production in the Mediterranean" finds four relevant items in the Italian

collection and four in the U.S. source while the query "The French Academy" obtains 24 relevant articles in the French corpus, 6 in the U.S. newspaper, and only 5 in the German collection and 2 in the Italian corpus.

Finally, we should mention the work of Baumgarten (1999) who presented a theoretical framework with which to solve both the collection fusion and collection selection problems, based on a probabilistic model. Baumgarten's work shows how to estimate the distribution of the rsv values corresponding to each collection, based on shifted gamma distributions. Within this framework, the distributed approach has a retrieval effectiveness that is close to that of the centralized model.

4.2. Evaluation based on set of queries provided in the CLEF 2000 test suite

In order to evaluate the retrieval effectiveness of these various merging strategies in our cross-lingual context, we used queries provided by the CLEF 2000 test suite, written in four different languages. In such circumstances, each collection was searched based on the best query formulation and the resulting average precision represents the best performance to be achieved by a multilingual system so far. The data shown in Tables 13a and 13b presents the average precision as achieved when searching our four collections. In these tables, we selected the round-robin approach as a baseline for comparisons, and the number of queries (40 in this case) is the same for all approaches, with the number of relevant documents being 2,266.

In our multilingual context, retrieval performance levels for the raw-score merging strategy are not very high, depicting a mean decrease of -7.09% over ten retrieval schemes. However, the difference between these two merging strategies can be considered as statistically significant for two experiments only ("doc=Okapi, query=npn" and "doc=nnn, query=nnn").

As a third merging approach, we might consider normalized score merging. In this case, we could normalize each document score based on the maximum retrieval status value achieved by the corresponding collection (evaluation listed under the heading "Normalized score (max)"). As a variant, we could also normalize the document score based on Equation 1, as suggested by Powell *et al.* (1999), in which each document score was normalized based on the document score achieved by the first retrieved item (rsv_{max}) and the retrieval status value obtained by the 1000th retrieved record (rsv_{min}). Merging the result lists

based on a normalized score improves the mean average precision by +3.12%, when using the maximum score or +5.5% when using Equation 1. However, these differences present a statistically significant improvement for two cases only, respectively three retrieval models out of ten.

	Average precision (% change)				
Model	Round-robin strategy 40 queries 2,266 rel doc	Raw-score merging 40 queries 2,266 rel. doc.	Normalized score (max) 40 queries 2,266 rel. doc.	Normalized score (Eq. 1) 40 queries 2,266 rel. doc.	CORI 40 queries 2,266 rel. doc.
Okapi-npn	24.43	15.04 (-38.4%)	26.94 (+10.3%)	26.94 (+10.3%)	14.98 (-38.7%)
Lnu-ltc	21.33	22.61 (+6.0%)	21.47 (+0.7%)	22.04 (+3.3%)	18.43 (-13.6%)
atn-ntc	21.69	18.79 (-13.3%)	23.93 (+10.3%)	24.53 (+13.1%)	15.74 (-27.4%)
dtu-dtc	21.27	23.83 (-12.0%)	22.56 (+6.1%)	23.45 (+10.2%)	18.49 (-13.1%)
ltn-ntc	20.37	19.41 (-4.7%)	20.18 (-0.9%)	21.12 (+3.7%)	15.81 (-22.4%)
ntc-ntc	13.46	13.94 (+3.6%)	12.88 (-4.3%)	13.06 (-3.0%)	12.87 (-4.4%)
ltc-ltc	13.29	14.13 (+6.3%)	13.64 (+2.6%)	13.80 (+3.8%)	12.63 (-5.0%)
lnc-ltc	13.31	14.15 (+6.3%)	12.97 (-2.6%)	13.29 (-0.2%)	12.60 (-5.3%)
bnn-bnn	12.29	9.67 (-21.3%)	14.84 (+20.8%)	14.81 (+20.5%)	9.27 (-24.6%)
nnn-nnn	6.40	4.65 (-27.3%)	5.65 (-11.7%)	5.96 (-6.9%)	4.34 (-32.2%)
Mean difference		-7.09%	+3.12%	+5.50%	-18.66%

Table 13a: Average precision of various result merging strategies using the queries provided in the CLEF 2000 test suite (Title only)

	Average precision (% change)				
Model	Round-robin strategy 40 queries 2,266 rel doc	Normalized CORI 40 queries 2,266 rel. doc.	Regression $rank_j$ 40 queries 2,266 rel. doc.	Regression $ln(rank_j)$ 40 queries 2,266 rel. doc.	Logistic regres. $ln(rank_j)$, rsv_j 40 queries 2,266 rel. doc.
Okapi-npn	24.43	26.74 (+9.5%)	19.77 (-19.1%)	23.77 (-2.7%)	26.52 (+8.6%)
Lnu-ltc	21.33	22.25 (+4.3%)	18.08 (-15.2%)	20.46 (-4.1%)	24.78 (+16.2%)
atn-ntc	21.69	22.96 (+5.9%)	17.43 (-19.7%)	21.33 (-1.7%)	25.58 (+17.9%)
dtu-dtc	21.27	21.97 (+3.3%)	17.75 (-16.6%)	20.52 (-3.5%)	25.47 (+19.7%)
ltn-ntc	20.37	19.88 (-2.4%)	16.63 (-18.4%)	19.83 (-2.7%)	24.05 (+18.1%)
ntc-ntc	13.46	13.21 (-1.9%)	11.86 (-11.9%)	14.10 (+4.8%)	15.83 (+17.6%)
ltc-ltc	13.29	13.26 (-0.2%)	12.04 (-9.4%)	13.47 (+1.4%)	15.18 (+14.2%)
lnc-ltc	13.31	13.25 (-0.5%)	11.85 (-11.0%)	13.39 (+0.6%)	14.82 (+11.3%)
bnn-bnn	12.29	13.18 (+7.2%)	11.93 (-2.9%)	13.60 (+10.7%)	14.69 (+19.5%)
nnn-nnn	6.40	6.42 (+0.3%)	0.31 (-95.2%)	2.10 (-67.2%)	7.12 (+11.3%)
Mean difference		+2.55%	-21.92%	-6.44%	+15.44%

Table 13b: Average precision of various result merging strategies using the queries provided in the CLEF 2000 test suite (Title only)

As a fourth merging approach, we evaluated the CORI system. In this case, we might use the original document score (rsv_j) and multiply it by the corresponding collection weight

w_i as described previously (see Table 13a, last column). Nevertheless, we just observed that the raw-score approach produces lower retrieval effectiveness than does the normalized score merging strategy. Therefore, instead using the original document score value, we first normalized them, using Equation 1. On the other hand, instead of using Equation 2 to compute the collection weight, we used the following equation, as suggested by Powell *et al.* (2000).

$$w'_i = (s_i - s_{min}) / (s_{max} - s_{min})$$

and computed the final document score based on a linear combination of the normalized document score rsv'_j and the normalized collection score w'_i as follows:

$$rsv''_j = (rsv'_j + 0.4 \cdot w'_i \cdot rsv'_j) / 1.4$$

This modified version of the CORI approach called "Normalized CORI" (third column of Table 13b) shows that the mean retrieval effectiveness improvement is 29.3% compared to the normal CORI approach. For example, using the search model "doc=Okapi, query=npn", the normal CORI merging strategy shows an average precision of 14.98 while for the normalized CORI retrieval effectiveness is 26.74 (+78.5%). On the other hand, when comparing this "normalized CORI" merging strategy with the round-robin method as shown in Table 13b, the differences across the ten retrieval schemes is small and only for the "doc=Okapi, query=npn" search engine can this difference be viewed as statistically significant.

As a fifth merging approach, we evaluated a merging collection based on the rank of the retrieved items. To account for the retrieval performances of the various collections (or for each language), we have fitted a linear regression for each language in which the explanatory variable was simply the rank (under the heading "Regression $rank_j$") or the logarithm of the rank (under the heading "Regression $ln(rank_j)$") as described in Franz *et al.* (1999). The data shown in Table 13b demonstrates that by estimating the probability of relevance based on the logarithm of the rank improves performance compared to simply using the rank. However, the differences between the round-robin and merging strategy based on the logarithm of the rank are not statistically significant.

As a sixth database merging strategy, we considered the logistic regression approach. In order to establish a common basis upon which to merging documents, we used the logarithm of the rank of the retrieved document ($ln(rank_j)$) together with its similarity value (rsv_j) as explanatory variables. Thus, for each collection and each retrieval scheme, we

estimated the underlying coefficients of the logistic regression using the R package (Venables and Ripley, 1999), freely available at http://cran.r-project.org. This merging strategy demonstrated statistical improvement for all retrieval schemes, as compared to the round-robin approach (see Table 13b).

Finally, average precision achieved for the Okapi search model using either the normalized score or the logistic regression merging strategy was around 26.6. Comparing this retrieval performance with the monolingual performances (see Table 3) is a indication of the "cost" of working with distributed collections.

4.3. Evaluation using translated queries

In the previous section, searches across the four collections were done using the request set provided by the CLEF 2000 test suite. In this section, we will evaluate the search engine's retrieval performance using English queries and also their respective translations obtained automatically using both the SYSTRAN system, the BABYLON bilingual dictionary (when taking into account both the first alternative or the first two translation alternates), and a combination of both automatic approaches. For all the experiments in this section, we used the same retrieval model, namely "doc=Okapi, query=npn".

In the top part of Table 14a, we report the average precision of our four collections using the CLEF 2000 formulation. In this part, we also indicate the average precision achieved using the best parameter setting for the automatic query expansion (see Section 3.3). The bottom part of Table 14a lists the average precision achieved using the translated queries, based on various automatic translation approaches and based on the English formulation of the requests. These retrieval effectiveness measures were obtained using the same retrieval scheme ("doc=Okapi, query=npn") on only one collection at a time (see Chapter 3).

The Table 14b and 14c lists the mean average precision obtained when the English set of queries provided in the CLEF 2000 test suite are translated into the other three languages and the four result lists are combined to produce a single answer list. In Table 14b and 14c, we selected the round-robin approach as a baseline for comparisons.

	Average precision			
Query (Title only) Model	English 33 queries 579 relevant	French 34 queries 528 relevant	Italian 34 queries 338 relevant	German 37 queries 821 relevant
Okapi-npn	37.26	41.62	33.98	31.42
+ best expand	37.26	42.22	40.91	37.71
Translated queries				
using SYSTRAN		29.64	20.79	22.59
using BABYLON 1		29.45	19.93	17.39
using BABYLON 2		21.37	17.47	18.61
combined		33.14	25.78	25.43
+ best expand		36.00	32.10	28.33

Table 14a: Average precision of our four corpora using the queries provided in the CLEF 2000 test suite and different automatic query translations (Okapi search model, Title only)

	Average precision				
Model	Round-robin strategy 40 queries 2,266 relev.	Raw-score merging 40 queries 2,266 relevant	Normalized score (max) 40 queries 2,266 relevant	Normalized score (Eq. 1) 40 queries 2,266 relevant	CORI 40 queries 2,266 relevant
CLEF 2000	24.43	15.04 (-38.4%)	26.94 (+10.3%)	27.50 (+12.6%)	14.98 (-38.7%)
+ expand	27.28	9.42 (-65.5%)	26.52 (-2.8%)	28.35 (+3.9%)	10.32 (-62.2%)
Translated					
SYSTRAN	16.71	9.71 (-41.9%)	17.46 (+4.5%)	17.78 (+6.4%)	9.66 (-42.2%)
BABYLON 1	15.17	10.05 (-33.8%)	17.30 (+14.0%)	17.60 (+16.0%)	9.90 (-34.7%)
BABYLON 2	13.68	9.68 (-29.2%)	15.57 (+13.8%)	15.63 (+14.3%)	9.08 (-33.6%)
combined	19.31	11.30 (-41.5%)	20.21 (+4.7%)	20.77 (+7.6%)	10.33 (-46.5%)
+ expand	21.40	9.42 (-56.0%)	21.97 (+2.7%)	22.82 (+6.6%)	10.28 (-52.0%)
Mean difference		-43.75%	+6.74%	+9.62%	-44.27%

Table 14b: Average precision of various merging strategies using four multi-lingual corpora (Okapi search model, Title only)

When searching in multi-lingual corpora using Okapi, the raw-score merging strategy does not provide interesting levels of retrieval performance (see Table 14b). In particular when using a query expansion approach, the same retrieval scheme produces really different document scores, thus invalidating the underlying assumption of the raw-score merging strategy. The normalized score merging shows an enhancement over the baseline that is statistically significant for the three search models (see Table 14b). In Table 14b last column, we can see that the CORI approach does not perform very well due to the use of document scores that are quite different across the collections. The normalized CORI system's retrieval performance shows retrieval effectiveness comparable to that of the round-robin approach

(see Table 14c, third column). For only one search and translation strategy (BABYLON 2) is the difference between these two merging approaches significant.

	Average precision				
Model	Round-robin strategy 40 queries 2,266 relev.	Normalized CORI 40 queries 2,266 relevant	Regression $rank_j$ 40 queries 2,266 relevant	Regression $ln(rank_j)$ 40 queries 2,266 relevant	Logistic regres. $ln(rank_j)$, rsv_j 40 queries 2,266 relevant
CLEF 2000	24.43	24.40 (-0.1%)	19.77 (-19.1%)	23.77 (-2.7%)	26.52 (+8.6%)
+ expand	27.28	25.90 (-5.1%)	20.87 (-23.5%)	26.65 (-2.3%)	29.86 (+9.5%)
Translated					
SYSTRAN	16.71	15.71 (-6.0%)	11.19 (-33.0%)	15.82 (-5.3%)	18.12 (+8.4%)
BABYLON 1	15.17	16.44 (+8.4%)	10.46 (-31.1%)	12.31 (-18.9%)	17.33 (+14.2%)
BABYLON 2	13.68	15.60 (+14.1%)	10.95 (-20.0%)	14.36 (+5.0%)	15.89 (+16.1%)
combined	19.31	18.36 (-4.9%)	14.25 (-26.2%)	18.11 (-6.2%)	20.74 (+7.4%)
+ expand	21.40	19.96 (-6.7%)	16.12 (-24.7%)	20.03 (-6.4%)	22.81 (+6.6%)
Mean difference		-0.06%	-25.36%	-5.26%	+10.12%

Table 14c: Average precision of various merging strategies using four multi-lingual corpora (Okapi search model, Title only)

As noted in the previous section, using the logarithm of the rank to merge the various result lists provides superior performance when compared to simply using the rank of the retrieved items. However, the difference between the round-robin and the merging strategy based on the logarithm of the rank is not statistically significant.

When merging the result lists based on the logistic regression approach, the differences in mean average precision favor this merging strategy over the round-robin model (last column of Table 14c). In the four cases, variations could be considered as statistically significant. When comparing the retrieval performance of the normalized score strategy (using Equation 1) and the logistic regression approach however, we could not find any significant differences in the average precision between the two merging strategies.

Finally, for this merged experiment, when using the requests provided by the CLEF 2000 test suite and automatically expanding them, the mean average precision achieved was around 29.1 % (normalized score based on Equation 1 or using the regression logistic merging strategy). Based on the same retrieval scheme and with our combined cross-lingual approach, we achieved a mean average precision of around 22.8% (a mean decrease of -21.6%).

5. Conclusion

Convinced that isolated retrieval effectiveness evaluations are not very useful, we have carried out experiments based on the various search strategies used for retrieving information from collections written in four different languages. We also evaluated various cross-language information retrieval models, where our experimental results show that:

- French, Italian or German collections can be accessed with the same retrieval models developed for the English corpora (see Table 3);
- the best retrieval model for the English collection is also the best for the three other languages (Table 3);
- using more search terms may significantly improve retrieval effectiveness (Table 4 and Appendix 3);
- indexing German documents using a 5-gram approach results in significantly better retrieval performance than does indexing based on words (Table 6). Moreover, combining both indexing schemes may sometimes lead to average precision enhancement;
- bilingual retrieval based on the query translation approach using only one source of evidence (machine translation or bilingual dictionary) is not really effective (Table 8);
- combining a bilingual dictionary and a machine translation approach may significantly enhance retrieval effectiveness (Table 8);
- combining the bilingual dictionary, the machine translation system and blind query expansion approaches may result in interesting performance levels, close to those of monolingual retrieval approaches (Tables 8 and 9);
- when analyzing the distribution of the number of translation alternatives provided by a bilingual dictionary, the majority of its entries consist of one or two translations (Table 7 and Figure 1);
- when different users translate the same set of requests manually, their different translations result in noticeable variations in retrieval performances (Table 11);
- when merging result lists obtained from different corpora written in various languages, the normalized score approach can be viewed as an appropriate first approach (Tables 13 and Tables 14), and the logistic regression merging strategy may be considered when the training data is available.

Of course, these findings need still be confirmed using other languages or other test-collections. Although we are fairly capable of building stopword lists and stemming

procedures for the English language, when used for other European languages these two IR tools still need improvement. For those languages having high frequencies of compound word constructions, it could still be worthwhile to know whether n-gram indexing approaches might produce higher levels of retrieval performance than enhanced word segmentation heuristics. Moreover, we could consider additional sources of evidence when translating a request or strategies that would weight translation alternatives appropriately. Finally, when searching in multiple collections that contain documents written in various languages, it might be worthwhile to look into better collection merging strategies or include intelligent selection procedures in order to avoid searching in a collection or in a language that does not contain any relevant document.

Acknowledgments

The author would like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system, without which this study could not have been conducted. The author also wishes to thank the anonymous referees for their useful and constructive comments in order to improve this paper.

Appendix 1. Weighting schemes

To assign an indexing weight w_{ij} that reflects the importance of each single-term T_j in a document D_i, we may take three different factors into account. They are represented by the following three code letters respectively:

- within-document term frequency, denoted by tf_{ij} (first letter);
- collection-wide term frequency, denoted by df_j (second letter);
- normalization scheme (third letter).

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$	atn	$w_{ij} = idf_j \cdot [0.5+0.5 \cdot tf_{ij} / \max tf_i.]$
nfn	$w_{ij} = \ln\left[n / df_j \right]$	npn	$w_{ij} = tf_{ij} \cdot \ln\left[(n - df_j) / df_j \right]$
Okapi	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) / \left(K + tf_{ij}\right)$	Lnu	$w_{ij} = \dfrac{\left(1 + \ln(tf_{ij}) / 1 + pivot\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$
lnc	$w_{ij} = \dfrac{\left(\ln(tf_{ij}) + 1\right)}{\sqrt{\sum_{k=1}^{t} \left((\ln(tf_{ik}) + 1)\right)^2}}$	ntc	$w_{ij} = \dfrac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \cdot idf_k\right)^2}}$
dtc	$w_{ij} = \left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right) \cdot idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln\left(\ln(tf_{ik}) + 1\right) + 1\right) \cdot idf_k\right)^2}$		
ltc	$w_{ij} = \left(\ln(tf_{ij}) + 1\right) idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left((\ln(tf_{ik}) + 1) \cdot idf_k\right)^2}$		
dtu	$w_{ij} = \dfrac{\left(\left(1 + \ln\left(1 + \ln(tf_{ij})\right)\right) \cdot idf_j / 1 + pivot\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$		

Table A.1: Weighting schemes

In Table A.1, n indicates the number of documents in the collection, document length (the number of indexing terms) of D_i is denoted by nt_i, the constant advl is set at 900, the constant b at 0.75, the constant k_1 at 2, the constant pivot at 125 and the constant slope at 0.1. For the Okapi weighting scheme, K represents the ratio between the length of D_i measured by l_i (sum of tf_{ij}) and the collection mean noted by advl.

Appendix 2. Query examples

<num> C001 <E-title> Architecture in Berlin
<num> C002 <E-title> The Electroweak Theory
<num> C003 <E-title> Drugs in Holland
<num> C004 <E-title> Floods in Europe
<num> C005 <E-title> European Union Membership
<num> C006 <E-title> French Conscientious Objectors
<num> C007 <E-title> Drug Use and soccer
<num> C008 <E-title> The Suicide of Pierre Bérégovoy
<num> C009 <E-title> Methane Deposits
<num> C010 <E-title> War and Radio
<num> C011 <E-title> New Constitution for South Africa
<num> C012 <E-title> Solar Temple
<num> C013 <E-title> Conference on Birth Control
<num> C014 <E-title> Tourism in the U.S.
<num> C015 <E-title> Competitiveness of European Industry
<num> C016 <E-title> The French Academy
<num> C017 <E-title> Bush Fire near Sydney
<num> C018 <E-title> Firefighter Casualties
<num> C019 <E-title> Gulf War Syndrome
<num> C020 <E-title> Single European Currency
<num> C021 <E-title> European Economic Area
<num> C022 <E-title> Airplane Runway Accidents
<num> C023 <E-title> Postmenopausal Pregnancy
<num> C024 <E-title> World Trade Organization
<num> C025 <E-title> Corruption in Italy
<num> C026 <E-title> Use of Wind Power
<num> C027 <E-title> Integration of German Immigrants
<num> C028 <E-title> Teaching Techniques for non-English Speakers
<num> C029 <E-title> Nobel Prize for Economics
<num> C030 <E-title> Supermarket Ceiling in Nice collapses
<num> C031 <E-title> Consumer Protection in the EU
<num> C032 <E-title> Female priests
<num> C033 <E-title> Cancer Genetics
<num> C034 <E-title> Alcohol Consumption in Europe
<num> C035 <E-title> Wolves in Italy
<num> C036 <E-title> Olive Oil Production in the Mediterranean
<num> C037 <E-title> Sinking of the Estonia
<num> C038 <E-title> Return of Military Remains
<num> C039 <E-title> Investments in Eastern Europe or Russia
<num> C040 <E-title> Privatisation of German Rail

Table A.2: CLEF 2000 English queries (Title only)

<num> C012 (both query translations failed in German)
<E-title> Solar Temple
<D-title manually translated> Sonnentempel
<D-title SYSTRAN> SolarBügel
<D-title BYBYLON> solar Heiligtum

<num> C017 (both query translations succeeded in Italian)
<E-title> Bush Fire near Sydney
<I-title manually translated> Incendi boschivi vicino a Sydney
<I-title SYSTRAN> Fuoco del cespuglio vicino a Sydney
<I-title BYBYLON> cespuglio fuoco stretto Sydney

<num> C026 (both query translations failed in Italian)
<E-title> Use of Wind Power
<I-title manually translated> Impiego dell'energia eolica
<I-title SYSTRAN> Uso di pontenza del vento
<I-title BYBYLON> usare di vento efficacia

<num> C029 (both query translations failed in French)
<E-title> Nobel Prize for Economics
<F-title manually translated> Le Prix Nobel d'économie
<F-title SYSTRAN> Prix Nobel pour des sciences économiques
<F-title BYBYLON> Nobel prix pour les sciences économiques
(in German, SYSTRAN translation failed but BYBYLON succeeded)
<D-title manually translated> Erster Nobelpreis für Wirtschaft
<D-title SYSTRAN> Nobelpreis für Volkswirtschaft
<D-title BYBYLON> Nobelpreis Preis für Wirtschaftswissenschaft

<num> C033 (both query translations succeeded in French)
<E-title> Cancer Genetics
<F-title manually translated> Tumeurs et génétique
<F-title SYSTRAN> La Génétique De Cancer
<F-title BYBYLON> Cancer génétique

<num> C037 (only SYSTRAN query failed in French)
<E-title> Sinking of the Estonia
<F-title manually translated> Naufrage du ferry-boat Estonia
<F-title SYSTRAN> Descente de l'Estonie
<F-title BYBYLON> naufrage de le Estonie

Table A.3: CLEF 2000 English queries and their various translations

Appendix 3. Evaluation of various query formulations

	Average precision (% change)		
Query Model / Mean indexing terms	Title 2.8 terms	Title-Desc 6.68 terms	Title-Desc-Narr 17.53 terms
doc=Okapi, que=npn	33.98	42.55 (+25.2%)	47.84 (+40.8%)
doc=Lnu, query=ltc	32.47	40.85 (+25.8%)	47.58 (+46.5%)
doc=atn, query=ntc	28.96	38.70 (+33.6%)	43.09 (+48.8%)
doc=dtu, query=dtc	31.04	41.75 (+34.5%)	46.18 (+48.8%)
doc=ltn, query=ntc	31.90	38.63 (+21.0%)	47.08 (+47.6%)
doc=ntc, query=ntc	20.35	27.03 (+32.8%)	32.77 (+61.0%)
doc=ltc, query=ltc	18.39	25.47 (+38.5%)	33.31 (+81.1%)
doc=lnc, query=ltc	21.25	29.43 (+38.5%)	37.36 (+75.8%)
doc=bnn, query=bnn	19.63	20.20 (+2.9%)	18.14 (-7.6%)
doc=nnn, query=nnn	15.15	15.81 (+4.4%)	19.38 (+27.9%)

Table A.4: Average precision of various monolingual search models using different query formulations (Italian collection, 34 queries)

	Average precision (% change)		
Query Model / Mean indexing terms	Title 1.95 terms	Title-Desc 5.6 terms	Title-Desc-Narr 15.45 terms
doc=Okapi, que=npn	31.64	38.76 (+23.4%)	40.17 (+27.8%)
doc=Lnu, query=ltc	27.66	34.83 (+25.9%)	36.97 (+33.7%)
doc=atn, query=ntc	31.30	35.16 (+12.3%)	36.67 (+17.1%)
doc=dtu, query=dtc	28.23	32.54 (+15.3%)	35.32 (+25.1%)
doc=ltn, query=ntc	28.22	31.42 (+11.3%)	28.65 (+1.5%)
doc=ntc, query=ntc	23.42	25.56 (+9.1%)	28.13 (+20.1%)
doc=ltc, query=ltc	21.51	24.67 (+14.7%)	27.90 (+29.7%)
doc=lnc, query=ltc	21.65	26.32 (+21.6%)	30.31 (+40.0%)
doc=bnn, query=bnn	23.44	16.50 (-29.6%)	7.07 (-69.8%)
doc=nnn, query=nnn	9.78	6.34 (-35.2%)	4.59 (-53.1%)

Table A.5: Average precision of various monolingual search models using different query formulations (German collection, 37 queries)

	Average precision (% change)		
Query Model / Mean indexing terms	Title 2.6 terms	Title-Desc 5.65 terms	Title-Desc-Narr 12.7 terms
doc=Okapi, que=npn	37.26	44.18 (+18.6%)	47.98 (+28.8%)
doc=Lnu, query=ltc	32.69	41.65 (+27.4%)	45.64 (+39.6%)
doc=atn, query=ntc	31.40	34.45 (+9.7%)	38.43 (+22.4%)
doc=dtu, query=dtc	31.96	38.34 (+20.0%)	41.55 (+30.0%)
doc=ltn, query=ntc	25.28	30.14 (+19.2%)	32.07 (+26.9%)
doc=ntc, query=ntc	18.11	23.48 (+29.7%)	28.00 (+54.6%)
doc=ltc, query=ltc	16.76	22.01 (+31.3%)	27.84 (+66.1%)
doc=lnc, query=ltc	17.70	25.31 (+43.0%)	32.38 (+82.9%)
doc=bnn, query=bnn	12.54	17.31 (+38.0%)	11.84 (-5.6%)
doc=nnn, query=nnn	9.69	10.36 (+6.9%)	10.74 (+10.8%)

Table A.6: Average precision of various monolingual search models using different query formulations (English collection, 33 queries)

	Average precision (% change)		
Query Title-Desc-Narr Model	German 5-grams	German words	German combined
doc=Okapi, que=npn	40.17	33.20 (-17.4%)	42.09 (+4.8%)
doc=Lnu, query=ltc	36.97	30.48 (-17.6%)	39.72 (+7.4%)
doc=atn, query=ntc	36.67	30.71 (-16.3%)	37.06 (+1.1%)
doc=dtu, query=dtc	35.32	31.27 (-11.5%)	38.64 (+9.4%)
doc=ltn, query=ntc	28.65	26.08 (-9.0%)	30.16 (+5.3%)
doc=ntc, query=ntc	28.13	22.49 (-20.1%)	28.78 (+2.3%)
doc=ltc, query=ltc	27.90	24.26 (-13.0%)	29.06 (+4.2%)
doc=lnc, query=ltc	30.31	24.65 (-18.7%)	31.16 (+2.8%)
doc=bnn, query=bnn	7.07	6.31 (-10.8%)	7.51 (+6.2%)
doc=nnn, query=nnn	4.59	6.43 (+40.1%)	4.76 (+3.7%)
Mean difference		-9.40%	+4.72%

Table A.7: Average precision of various indexing and searching strategies based on monolingual requests and documents (German collection, 37 queries)

Appendix 4. Evaluation of various bilingual approaches

	Average precision (% change)				
Model	Monolingual	SYSTRAN	BABYLON 1	BABYLON 2	Combined
Okapi-npn	33.98	20.79 (-38.8%)	19.93 (-41.3%)	17.47 (-48.6%)	25.78 (-24.1%)
Lnu-ltc	32.47	19.70 (-39.3%)	18.96 (-41.6%)	18.16 (-44.1%)	24.62 (-24.2%)
atn-ntc	28.96	16.77 (-42.1%)	15.94 (-45.0%)	11.80 (-59.3%)	21.48 (-25.8%)
dtu-dtc	31.04	20.08 (-35.3%)	18.92 (-39.0%)	15.09 (-51.4%)	22.04 (-29.0%)
ltn-ntc	31.90	21.61 (-32.3%)	19.36 (-39.3%)	16.37 (-48.7%)	23.62 (-26.0%)
ntc-ntc	20.35	14.02 (-31.1%)	13.16 (-35.3%)	13.14 (-35.4%)	15.58 (-23.4%)
ltc-ltc	18.39	10.97 (-40.3%)	8.93 (-51.4%)	9.41 (-48.8%)	13.38 (-27.2%)
lnc-ltc	21.25	13.08 (-38.4%)	9.75 (-54.1%)	10.50 (-50.6%)	15.78 (-25.7%)
bnn-bnn	19.63	11.31 (-42.4%)	8.03 (-59.1%)	7.22 (-63.2%)	9.88 (-49.7%)
nnn-nnn	15.15	11.72 (-22.6%)	12.50 (-17.5%)	11.77 (-22.3%)	12.45 (-17.8%)
Mean difference		-36.27%	-42.38%	-47.24%	-27.30%

Table A.8: Average precision of various translating strategies using the Italian collection (Title only, using 34 English queries)

	Average precision (% change)				
Model	Monolingual	SYSTRAN	BABYLON 1	BABYLON 2	Combined
Okapi-npn	31.64	22.59 (-28.6%)	17.39 (-45.1%)	18.61 (-41.2%)	25.43 (-19.6%)
Lnu-ltc	27.66	18.74 (-32.2%)	15.01 (-45.7%)	16.30 (-41.1%)	21.79 (-21.2%)
atn-ntc	31.30	21.78 (-30.4%)	14.42 (-53.9%)	14.63 (-53.3%)	25.25 (-19.3%)
dtu-dtc	28.23	18.12 (-35.8%)	13.28 (-53.0%)	14.15 (-49.9%)	22.70 (-19.6%)
ltn-ntc	28.22	17.99 (-36.3%)	14.32 (-49.3%)	14.51 (-48.6%)	20.20 (-28.4%)
ntc-ntc	23.42	15.63 (-33.3%)	12.01 (-48.7%)	11.91 (-49.1%)	16.05 (-31.5%)
ltc-ltc	21.51	13.47 (-37.4%)	9.28 (-56.9%)	9.57 (-55.5%)	14.52 (-32.5%)
lnc-ltc	21.65	13.29 (-38.6%)	9.78 (-54.8%)	10.07 (-53.5%)	15.34 (-29.1%)
bnn-bnn	23.44	12.75 (-45.6%)	10.05 (-57.1%)	7.30 (-68.9%)	14.17 (-39.5%)
nnn-nnn	9.78	6.82 (-30.3%)	7.63 (-22.0%)	4.65 (-52.5%)	6.07 (-37.9%)
Mean difference		-34.80%	-48.60%	-51.30%	-27.82%

Table A.9: Average precision of various translating strategies using the German collection (Title only, using 37 English queries)

References

Adriani, M. (2001). English-Dutch CLIR using query translation techniques. In C. Peters (Ed.), Results of the CLEF-2001 cross-language system evaluation campaign, (pp. 143-146). Sophia-Antipolis: ERCIM.

Ballesteros, L. & Croft, B. W. (1998). Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st International Conference of the ACM-SIGIR'98, (pp. 64-71). New York: The ACM Press.

Baumgarten, C. (1999). A probabilistic solution to the selection and fusion problem in distributed information retrieval. In Proceedings of the 22nd International Conference of the ACM-SIGIR'99, (pp. 246-253). New York: The ACM Press.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Braschler, M. & Schäuble, P. (2001). Experiments with the Eurospider retrieval system for CLEF 2000. In C. Peters (Ed.), Cross-language information retrieval and evaluation, LNCS 2069, (pp. 140-148). Berlin: Springer-Verlag.

Broglio, J., Callan, J. P., Croft, W. B. & Nachbar, D. W. (1995). Document retrieval and routing using the INQUERY system. In Proceedings of TREC'3, (pp. 29-38). Gaithersburg: NIST Publication #500-225.

Brown, P. F., Pietra, S. A. D., Pietra, V. D. J. & Mercer, R. L. (1993). The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263-312.

Buckley, C., Singhal, A., Mitra, M. & Salton, G. (1996). New retrieval approaches using SMART. In Proceedings of TREC'4, (pp. 25-48). Gaithersburg: NIST Publication #500-236.

Callan, J. P., Lu, Z. & Croft, W. B. (1995). Searching distributed collections with inference networks. In Proceedings of the 18th International Conference of the ACM-SIGIR'95 (pp. 21-28). New York: The ACM Press.

Chen, A. (2001). Multilingual information retrieval using English and Chinese queries. In Proceedings of CLEF-2001 (p. 21-27). Sophia-Antipolis: ERCIM EEIG.

Chen, A., Jiang, H. & Gey, F. (2001). English-Chinese cross-language IR using bilingual dictionaries. In Proceedings TREC-9. Gaithersburg: NIST Publication (to appear).

David, M. & Ogden, W. C. (1997). QUILT: Implementing a large-scale cross-language text retrieval system. In Proceedings of the 20th International Conference of the ACM-SIGIR'97 (pp. 92-98). New York: The ACM Press.

Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In Proceedings of TREC'2, (pp. 105-115). Gaithersburg: NIST Publication #500-215.

EMarketer (2001). The world's leading provider on Internet statistics. http://www.emarketer.com/analysis/edemographics/20010227_edemo.html

Flury, B. (1997). *A first course in multivariate statistics*. New York: Springer.

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, 24, 19-35.

Fox, E. A. & Shaw, J. A. (1994). Combination of multiple searches. In Proceedings TREC'2, (pp. 243-249). Gaithersburg: NIST Publication #500-215.

Franz, M., Scott McCarley, J. & Todd Ward, R. (2000). Ad hoc, cross-language and spoken document information retrieval at IBM. In Proceedings TREC-8, (pp. 391-398). Gaithersburg: NIST Publication #500-246.

Franz, M., Scott McCarley, J. & Roukos, S. (1999). Ad hoc and multilingual information retrieval at IBM. In Proceedings TREC-7, (pp. 157-168). Gaithersburg: NIST Publication #500-242.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A. & Lochbaum K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In Proceedings of the 11th International Conference of the ACM-SIGIR'88, (pp. 465-480). New York: The ACM Press.

Gachot, D. A., Lange, E. & Yang, J. (1998). The SYSTRAN NLP browser: An application of machine translation technology. In G. Grefenstette (Ed.), Cross-language information retrieval, (pp. 105-118). Boston: Kluwer.

Gale, W. A. & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), 75-102.

Global Reach (2001). Global Internet statistics (by language). http://www.euromktg.com/globstats/

Grefenstette, G. (Ed.) (1998). *Cross-language information retrieval*. Amsterdam: Kluwer.

Hedlund, T., Pirkola, A. & Järvelin, K. (2001). Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing & Management*, 33(2), 147-161.

Hiemstra, D., Kraaij, W., Pohlmann, R. & Westerveld, T. (2001). Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In C. Peters (Ed.), Cross-language information retrieval and evaluation, LNCS 2069, (pp. 102-115). Berlin: Springer-Verlag.

Hosmer, D., & Lemeshow, S. (1989). *Applied logistic regression*. New-York: John Wiley & Sons.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In Proceedings of the 16th International Conference of the ACM-SIGIR'93, (pp. 329-338). New York: The ACM Press.

Hull, D. & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th International Conference of the ACM-SIGIR'96, (pp. 49-57). New York: The ACM Press.

Jansen, B. J., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing & Management*, 36(2), 207-227.

Jones, G. J. F. & Lam-Adesina, A. M. (2001). Exeter at CLEF-2001: Experiments with machine translation for bilingual retrieval. In C. Peters (Ed.), Results of the CLEF-2001 cross-language system evaluation campaign, (pp. 105-114). Sophia-Antipolis: ERCIM.

Kraaij, W., Pohlmann, R. & Hiemstra, D. (2000). Twenty-One at TREC-8: Using language technology for information retrieval. In Proceedings TREC-8, (pp. 285-300). Gaithersburg: NIST Publication #500-246.

Kraaij, W. & Pohlmann, R. (1996). Viewing stemming as recall enhancement. In Proceedings of the 19th International Conference of the ACM-SIGIR'96, (pp. 40-48). New York: The ACM Press.

Kraaij, W. (2001). TNO at CLEF-2001: Comparing translation resources. In C. Peters (Ed.), Results of the CLEF-2001 cross-language system evaluation campaign, (pp. 29-40). Sophia-Antipolis: ERCIM.

Kwok, K. L., Grunfeld L., Dinstl, N. & Chan, M. (2001). TREC-9 cross-language, Web and question-answering track experiments using PIRCS. In Proceedings TREC-9. Gaithersburg: NIST Publication (to appear).

Kwok, K. L., Grunfeld, L. & Lewis, D. D. (1995). TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In Proceedings of TREC'3, (pp. 247-255). Gaithersburg: NIST Publication #500-225.

Landauer, T. K. & Littman, L. M. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In Proceedings of the 6th Conference of Electronic Text Research, (pp. 31-38).

Le Calvé, A., & Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing & Management*, 36(3), 341-359.

Lee, J. H. (1997). Analyses of multiple evidence combination. In Proceedings of the 20th International Conference of the ACM-SIGIR'97 (pp. 267-276). New York: The ACM Press.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 22-31.

Manmatha, R., Rath, T. & Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. In Proceedings of the 24th International Conference of the ACM-SIGIR'2001, (pp. 267-275). New York: The ACM Press.

Mayfield, J., McNamee, P. & Piatko, J. (2000). The JHU/APL HAIRCUT system at Trec-8. In Proceedings TREC-8, (pp. 445-452). Gaithersburg: NIST Publication #500-246.

McNamee, P., Mayfield, J. & Piatko, C. (2001). A language-independent approach to European text retrieval. In C. Peters (Ed.), Cross-language information retrieval and evaluation, LNCS 2069, (pp. 129-139). Berlin: Springer-Verlag.

McNamee, P. & Mayfield, J. (2001). JHU/APL experiments at CLEF: Translation resources and score normalization. In C. Peters (Ed.), Results of the CLEF-2001 cross-language system evaluation campaign, (pp. 121-128). Sophia-Antipolis: ERCIM.

Miller, D., Leek, T. & Schwartz, R. (1999). A hidden Markov model information retrieval system. In Proceedings of the 22nd International Conference of the ACM-SIGIR'99, (pp. 214-221). New York: The ACM Press.

Moffat, A. & Zobel, J. (1995). Information retrieval systems for large document collections. In Proceedings of TREC'3, (pp. 85-93). Gaithersburg,: NIST Publication #500-225.

Nie, J. Y. & Ren F. (1999). Chinese information retrieval: Using characters or words? *Information Processing & Management*, 35(4), 443-462.

Nie, J. Y. & Simard, M. (2001). Using statistical translation models for bilingual IR. In C. Peters (Ed.), Results of the CLEF-2001 cross-language system evaluation campaign, (pp. 75-80). Sophia-Antipolis: ERCIM.

Nie, J. Y., Simard, M., Isabelle, P. & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In Proceedings of the 22nd International Conference of the ACM-SIGIR'99, (pp. 74-81). New York: The ACM Press.

Oard, D. & Dorr, B. J. (1996). A survey of multilingual text retrieval. Institute for advanced computer studies and computer science department, University of Maryland, http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps.

Oard, D. W. & Resnik, P. (1999). Support for interactive document selection in cross-language information retrieval *Information Processing & Management*, 35(4), 363-379.

Peters, C. (Ed.) (2001). *Cross-language information retrieval and evaluation.* Lecture Notes in Computer Science, 2069. Berlin: Springer-Verlag.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Proceedings of the 21st International Conference of the ACM-SIGIR'98, (pp. 55-63). New York: The ACM Press.

Ponte, J. & Croft, W. B. (1998). A language modeling approach to information retrieval. In Proceedings of the 21st International Conference of the ACM-SIGIR'98, (p. 275-281). New York: The ACM Press.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

Powell, A. L., French, J. C., Callan, J., Connell, M. & Viles, C. L. (2000). The impact of database selection on distributed searching. In Proceedings of the 23rd International Conference of the ACM-SIGIR'2000, (pp. 232-239). New York: The ACM Press.

van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths, 2nd ed.

Robertson, S. E., Walker, S. & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

Robertson, S. E. & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3), 129-146.

Salton, G. & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. New-York: McGraw-Hill.

Salton, G. (1971). Automatic processing of foreign language documents. In G. Salton (Ed.), The SMART retrieval system, Experiments in automatic document processing, (pp. 206-219). Englewood Cliffs, NJ: Prentice-Hall.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 495-512.

Savoy, J., Ndarugendamwo, M. & Vrajitoru. D. (1996). Report on the TREC-4 experiment: Combining probabilistic and vector-space schemes. In Proceedings TREC-4, (pp. 537-548). Gaithersburg: NIST Publication 500-236.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), 944-952.

Savoy, J. & Rasolofo, Y. (2001). Report on the TREC-9 experiment: Link-based retrieval and distributed collections. In Proceedings TREC-9. Gaithersburg: NIST Publication (to appear). .

Sheridan, P. & Ballerini, J. P. (1996). Experiments in multilingual information retrieval using SPIDER system. In Proceedings of the 19th International Conference of the ACM-SIGIR'96, (pp. 58-65). New York: The ACM Press.

Singhal, A., Choi, J., Hindle, D., Lewis, D. D. & Pereira, F. (1999). AT&T at TREC-7. In Proceedings TREC-7, (pp. 239-251). Gaithersburg: NIST Publication #500-242.

Sparck Jones, K. & Bates, R. G. (1977). Research on automatic indexing 1974-1976. Technical Report, Computer Laboratory, University of Cambridge, UK.

Spink, A. & Saracevic, T. (1997). Interactive information retrieval: sources and effectiveness of search terms during mediated online searching. *Journal of the American Society for Information Science*, 48(8), 741-761.

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Sproat, R. (1992). *Morphology and computation*. Cambridge: The MIT Press.

Venables, W. N. & Ripley, B. D. (1999). *Modern applied statistics with S-plus.* 3rd Ed., New York: Springer.

Voorhees, E. M., Gupta, N. K. & Johnson-Laird, B. (1995a). The collection fusion problem. In Proceedings of TREC'3, (pp. 95-104). Gaithersburg: NIST Publication #500-225.

Voorhees, E. M., Gupta, N. K. & Johnson-Laird, B. (1995b). Learning collection fusion strategies. In Proceedings of the 18th International Conference of the ACM-SIGIR'95, (pp. 172-179). New York: The ACM Press.

Voorhees, E. M. (1996). Siemens TREC-4 report: Further experiments with database merging. In Proceedings of TREC'4, (pp. 121-130). Gaithersburg: NIST Publication #500-236.

Voorhees, E. M. & Harman, D. (2000). Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing & Management*, 36(1), 3-35.

Xu, J. & Croft W. B. (1999). Cluster-based language models for distributed retrieval. In Proceedings of the 22nd International Conference of the ACM-SIGIR'99, (p.254-261). New York: The ACM Press.

Xu, J. & Weischedel, R. (2001). TREC-9 cross-lingual retrieval at BBN. In Proceedings TREC-9. Gaithersburg: NIST Publication (to appear).