

# Using Structured Queries for Disambiguation in Cross-Language Information Retrieval

**David A. Hull**

Rank Xerox Research Centre  
6 chemin de Maupertuis, 38240 Meylan France  
hull@grenoble.rxrc.xerox.com

June 23, 1997

## **Abstract**

Bilingual transfer dictionaries are an important resource for query translation in cross-language text retrieval. However, term translation is not an isomorphic process, so dictionary-based systems must address the problem of ambiguity in language translation. In this paper, we claim that boolean conjunction (the AND operator) provides simple and automatic disambiguation in the target language. We derive a new weighted boolean model based on a probabilistic formulation and apply it to the cross-language text retrieval problem. The results suggest that the weighted boolean model is highly effective for general text retrieval, but more experimental evidence is needed to conclude that it is particularly advantageous for cross-language application. Nonetheless, the preliminary results are quite promising.

## **1 Introduction**

With the ongoing development of multilingual information retrieval systems, researchers are becoming increasingly interested in the problem of cross-language information retrieval (CLIR), which has been defined as taking a query in one language and retrieving relevant documents in other languages. This problem is considerably more complex than traditional information retrieval because some method for query or document translation must be developed before one can use traditional IR term matching and document ranking algorithms. Many approaches have been proposed and tested, including using such resources as bilingual dictionaries, thesauri, corpora, and/or machine translation systems. This paper reviews some of the recent theoretical research into using these strategies for cross-language retrieval, a study of which reveals that a lot of

progress has been made on the query translation problem. However, determining the best way to deal with ambiguity in query translation remains one of the most important research issues.

In our previous work, we have shown that dictionary-based query translation, where each term or phrase in the query is replaced by a list of all of its possible translations, represents an acceptable first pass at cross-language information retrieval [13]. This approach takes advantage of the fact that modern IR systems are sufficiently robust that the important concepts in the query still tend to dominate, despite the presence of spurious translation equivalents. However, we have only dodged the ambiguity problem, not resolved it, and for a significant fraction of the queries, this approach is not acceptable. In our current work, we examine whether the use of structured queries can help to reduce ambiguity and lead to better performance in cross-language information retrieval.

The boolean model provides a natural method for reducing the impact of spurious translations by enforcing conjunctive relationships between query terms. As a query language based on strict boolean constraints has other drawbacks, we develop and explore the potential of a weighted boolean model, where documents are scored according to how well they satisfy the boolean constraints using a probabilistic weighting scheme. We begin by reviewing recent research progress in CLIR and discussing the advantages and disadvantages of the approaches which have been tested. We motivate and develop a new weighted boolean model, and conclude by presenting some results which measure the effectiveness of the model and its suitability for cross-language text retrieval.

## 2 Background: Cross-Language Text Retrieval

Cross-language text retrieval has long been studied in the framework of controlled vocabulary systems. Early research [25, 21] demonstrated that cross-language systems can perform on par with monolingual systems given a carefully (i.e. manually) constructed bilingual thesaurus. However, controlled vocabulary systems are less than ideal for modern text retrieval for a number of reasons. The size and dynamic nature of modern information sources (such as the World Wide Web) makes manual document indexing and thesaurus construction a difficult task. See Oard's 1996 survey paper [18] for an extensive review of the history of research in cross-language text retrieval. In this section, we focus on recent research, concentrating on the material presented at the recent SIGIR '96 workshop on cross-language text retrieval [9]. There are three basic types of resources for query translation: dictionaries (or thesauri), monolingual and multilingual corpora, and user input. Different cross-language IR systems use various combinations of these resources.

## 2.1 Corpus Based Systems

Corpus-based systems use parallel and or comparable corpora for query translation. In a parallel corpus, the same documents exist in the source language and in one or more translations. In a comparable corpus, the relationship between documents is less clear, and several different definitions have emerged in recent research. Sheridan [26, 9] uses a very strict definition: comparable documents are about the same event and written at the same time, each in a different language. Yamabana [9] uses the term to describe documents from the same domain (e.g. politics, medicine) while Peters [20, 9] requires only that they be related by genre and style of diction. Clearly, as documents become less and less comparable, using them for CLIR becomes more and more difficult. For convenience, we assume that comparable corpora are not parallel in the following discussions.

Dumais, Landauer, and Littman at Bellcore [14, 9] and Oard at the University of Maryland [17, 9] have used a matrix reduction technique called Latent Semantic Indexing (LSI) to extract language independent term and document representations from a parallel corpus. Essentially, LSI applies a singular value decomposition to the large, sparse term-document cooccurrence matrix (including terms from all parallel versions of the document) and extracts a subset of the singular vectors to form a new vector space. Bellcore has demonstrated that using this technique, a document in one language will retrieve its mate in the other language first 98-99% of the time. If a shorter pseudo-query is constructed using only five important words from the document, performance drops to 55-60%, but the equivalent document can still be found in the top ten 92-93% of the time. While these results paint an encouraging picture of the ability of the LSI representation to capture meaning across languages, it is still unclear how well LSI would perform in domains outside that of the parallel training corpus. Unfortunately, parallel corpora are not common in general language, tending rather to be restricted to specialized domains.

Sheridan et al. at ETH Zurich [26, 9] have created a reference corpus for cross-language information retrieval by aligning news stories in German and Italian by topic label and date and merging them to create pseudo-parallel documents. German queries are then translated into Italian by using these combined documents for query expansion in conjunction with a word similarity thesaurus. The similarity thesaurus technique is similar to LSI in that terms are related to one another by their distribution across documents, but it performs no dimensionality reduction of the term-document matrix. The Italian terms from the expanded query can be directly compared to Italian documents. This technique has been evaluated for information retrieval, and it turns out to be about half as effective as monolingual text retrieval (in terms of average precision). Like LSI, this technique may also have a domain dependence problem.

## 2.2 Dictionary Based Systems

Dictionary-based systems perform query translation by looking up component terms and phrases in a bilingual dictionary and forming a target language query by concatenating some or all of the translations. Radwan and Fluhr [23] have constructed a cross-language text retrieval system known as SPIRIT which uses a ranked boolean retrieval system in conjunction with bilingual term, compound, and idiom dictionaries for query translation and document retrieval. Their goal was to compare dictionary-based translation to machine translation, and they found that their dictionary-based cross-language system is 75-80% as effective as monolingual retrieval while machine translation of the query is only 60-65% as effective. It is important to realize that they created a domain dependent terminology dictionary and performed extensive manual editing to achieve this level of performance. However, the machine translation system also used special topical dictionaries, so their test represent a fair comparison.

In previous research work at RXRC, we constructed a dictionary-based cross-language system using much the same model as Radwan and Fluhr [13]. We compared an automatic word-based translation model to one with manual correction of the dictionary and to one with a comprehensive multi-word terminology dictionary. The baseline model is about 60% as effective as monolingual retrieval, and this improves to 65-70% with manual correction of the dictionary then to 90% with the addition of multi-word terminology. However, we constructed the terminology dictionary specially for these experiments, so this represents an ideal or upper-bound performance figure. Note that these percentages are based on a high precision measure and so are not directly comparable to others presented in this section, which use precision averaged at fixed recall. Furthermore, we discovered that roughly 20% of the queries are significantly penalized due to ambiguity introduced in query translation. This is the most significant theoretical problem with dictionary-based models which has led to the development of hybrid systems. Dictionary-based systems also have a coverage problem, which led the researchers above to resort to manual techniques for dictionary cleaning and terminology generation.

## 2.3 Hybrid Systems

Hybrid systems use a combination of dictionaries, corpora, and user interaction for query translation. Peters and Picchi at CNR Pisa [20, 9] have looked into methods to automatically generate translation equivalents using comparable corpora. Each source language term is described by a profile of all those words that cooccur within a fixed window. The same process is also used for terms in the target language. Each source language profile is then translated into a target language profile using a bilingual dictionary, and the profiles for target language terms which are most similar are identified. The top ranked target language terms are treated as translation equivalents for the source language term. CNR

presents some examples of this process, but no extensive performance testing has yet been completed.

Yamabana et al. at NEC [7, 9] have built an English/Japanese multilingual retrieval system which uses a bilingual dictionary, comparable corpora, user interaction, and machine translation. For terms in the dictionary with more than one translation equivalent, the best equivalent for a particular domain is selected using a method based on statistical cooccurrence over comparable corpora. Each source language term is matched to the most similar term in the same language. The target language equivalents of this pair are then compared and the ones with highest statistical cooccurrence frequency are selected as being appropriate in the selected domain. In experiments using newspaper texts, NEC find that this method is 75-85% accurate. While quite impressive for comparable corpora, this performance level is not high enough to always produce good results in information retrieval. Therefore, the NEC system also has an interactive user interface which allows the searcher to select the most appropriate translation equivalent using term definitions in his or her own language. Finally, the retrieved documents are passed through a machine translation system before being presented to the user.

Ballesteros and Croft at the University of Massachusetts [1, 9] have built a cross-language IR system that attacks the ambiguity problem by using extensive query expansion. Their hypothesis is that additional terms which are related to the primary concepts in the query are likely to be relevant and that by adding these terms to the query, the impact of incorrect equivalents generated during the translation process can be reduced. Query expansion is performed using an automatic feedback technique to select new terms which occur frequently in conjunction with query terms. This expansion can be conducted either before (in the source language) or after (in the target language) query translation. Pre-translation feedback is conducted on an independent source language corpus, while post-translation feedback uses the corpus being searched. UMass finds that the best approach is to use a combination of pre- and post-translation feedback. The cross-language performance ranges initially between 40-50% of the monolingual level and increases to 60-65% after query expansion.

Davis at New Mexico State University [5, 9] also begins with dictionary-based query translation and has developed two strategies that perform direct disambiguation to select the best translation equivalent. His system applies a part of speech tagger to the query, which allows the system to select only those equivalents from the dictionary which have the same part of speech. It then measures the similarity each source language query term and the remaining equivalents to aligned sentences in a parallel corpus. Disambiguation is performed by selecting the equivalent whose sentence ranking is most similar to the source language term. Rankings can be compared across languages because the sentences are aligned. The cross-language performance rises from 40-50% for the initial translation to 70-75% after applying both disambiguation strategies.

## 2.4 Summary

It is difficult to compare the experimental results above because they are based on different test corpora and different evaluation measures. However, it seems fair to conclude that a simple dictionary-based cross-language system is about half as effective as its monolingual counterpart (on average, variation on individual queries can be substantial). Performance can be significantly improved by adding carefully constructed terminology dictionaries and corpus-based resources, but there is still an important gap between the best of these systems and the monolingual standard. The benefits of including corpus-derived resources are not surprising, because dictionaries and corpora can be thought of as complementary resources. Dictionaries provide broad and shallow coverage while corpora tend to provide narrow (domain-specific) but deep (more terminology) coverage of the language.

Researchers have produced an impressive array of different techniques to take advantage of the wealth of material in bilingual corpora. However, many of these methods will suffer when the domains of the training corpora and the search corpora are not sufficiently related. The notable exception is post-translation query expansion (applied by UMass) where the search corpus itself is used for expansion. Oard [19] suggests another important corpus-based technique that has not yet been explored, deriving translation probability vectors. This approach is epitomized by the work of Brown et al. [2] at IBM who align a parallel corpus on the word and phrase level using sophisticated models from information theory. The aligned corpus can then be used to extract probability estimates for term translation. While this technique will also suffer from the domain-relevance problem, it is a powerful approach for weighting translation equivalents which does not require full disambiguation.

It is striking (but not surprising) how few research systems incorporate user interaction into the query translation process. Most systems (with the exception of NEC) are focusing on automatic query translation and so do not try to leverage off the searcher. The experiments presented in this paper are no exception. The problem is that it is extremely difficult to quantify the impact of a searcher's interactive participation in the query construction problem and to make sure that it does not interact with other experimental variables. User studies of this kind are time-consuming and resource intensive. It is especially tricky to account for searcher influence in a multilingual setting because the searcher's foreign language knowledge is likely to be widely variable, yet it must be carefully accounted for. Issues concerning the user interface have been explored by Pollitt [22] and this represents an important area for future research.

### 3 Motivating and Building the Model

Previous research has demonstrated that ambiguity is a key problem in dictionary-based translation. There is another, more subtle issue which has not yet been addressed. Some terms have only one translation while others have many translations. The latter terms will have higher weight in the target language query than former, solely because of this property, which is magnified when many of the translations are synonyms and/or tend to cooccur together. Unfortunately, important content-bearing terms generally have only a few translations, so this effect is exactly the opposite of what is desired. It is unclear whether this translation weighting problem has any impact on performance but we hope to answer that question in our information retrieval experiments.

Our goal is to simultaneously address the ambiguity and the translation weighting problem by taking advantage of the structure in a boolean query. Boolean disjunction (the OR operator) is a natural way to link together many translation equivalents without dramatically increasing the weight of the underlying concept. Therefore, our weighted boolean CLIR system connects translation equivalents of the same term with the OR operator. Boolean conjunction (the AND operator) is likely to be an effective strategy for disambiguation if one believes the following hypothesis: "The correct translation equivalents of two or more query terms are much more likely to occur together in target language documents than in conjunction with any incorrect translation equivalents." One can think of incorrect translations as (relatively) random noise added to the query. The chance that medium to low frequency noise terms (derived from different source language words) will occur together regularly is remote.

The weighted boolean model has a number of advantages when compared to the techniques in the previous section. The search corpus itself is being used for disambiguation, so domain relevance is guaranteed. No additional reference corpora are required. User knowledge is incorporated inherently into the disambiguation process via query construction without requiring a subsequent feedback step (as proposed by NEC). There is no need to arbitrarily select a single best translation equivalent. Often the bilingual dictionary contains many synonyms, and the weighted boolean model allows one to take full advantage of this natural query expansion effect without the risk of assigning too much importance to the translated term. It remains to test whether the weighted boolean model is effective in retrieval experiments.

The strict boolean model has a number of known weaknesses: (1) no ranking of the retrieved set, (2) no concept of term importance, and (3) no control over the size of the retrieved set. These problems have been addressed by a number of extended boolean models which replace the strict logical expressions of the traditional model with fuzzy matching functions. Fox [8] and Lee [15] present nice surveys of the range of potential models. Modern systems have also worked on incorporating term proximity information into the weight functions [4] and with adding boolean operators to a probabilistic model in the context

of a Bayesian Inference Network [27]. When properly implemented, weighted boolean models can perform as well as the best vector processing models [10]. Their one key disadvantage is that searchers must spend more time and energy on careful query construction.

None of the extended boolean models described in the literature are entirely satisfactory for our purposes, so we have decided to develop new boolean weight functions for our cross-language experiments based on probabilistic principles. We use a much-simplified version of a completely general boolean model, following the model adopted by Hearst [11]. The query input format consists of a series of windows, one per line. Each window is considered a concept and all terms entered in that window are combined using an OR operator. The user has the option to designate the importance of each concept (ranging from not important to mandatory), which is translated by the system into a concept weight, and the concepts are then combined using a weighted AND operator. The model is based on the following simple probabilistic relations. Given independent events  $A$  and  $B$ :

$$P(A \text{ AND } B) = P(A \cap B) = P(A) * P(B)$$

$$P(A \text{ OR } B) = P(A \cup B) = 1 - (1 - P(A)) * (1 - P(B))$$

Given the term independence assumption, we can generalize these principles to construct a probabilistic indexing boolean model. Let  $t_i$  be a term and  $d$  be a document, then:

$$p(t_1 \text{ AND } \dots \text{ AND } t_n | d) = \prod_{i=1}^n p(t_i | d)$$

$$p(t_1 \text{ OR } \dots \text{ OR } t_n | d) = 1 - \prod_{i=1}^n (1 - p(t_i | d))$$

The model has been further generalized to include a user-specified concept importance scale that ranges between 1 and 3 (1 = not important, 2 = important, 3 = mandatory). This additional information is incorporated into the evaluation of the query, giving the users additional flexibility in query construction. The detailed mathematical formulation of the general model is given in [12], which should soon be available.

This approach also requires a strategy for estimating  $p(t_i | d)$ , the probability that term  $t_i$  is associated with document  $d$ . One could attempt to estimate  $p(t_i | d)$  directly using a 2-Poisson or negative binomial model. However, these models are extremely expensive to compute on the scale required in information retrieval, so we instead fall back on the approximation to the 2-Poisson Model developed by Robertson and Walker and known as BM25 [24] which has proven to be highly effective in information retrieval experiments:

$$\text{doc. weight} = \frac{P * (k + 1) * \text{tf}}{K + \text{tf}}$$

$$K = k * (1 - b) + b * \frac{\text{len}}{\text{avg.len}}, \quad P = \log\left(\frac{N - n}{n}\right)$$

where  $tf$  is the term frequency,  $len$  is the document length,  $avg.len$  is the average document length,  $P$  is the probabilistic term importance weight, and  $(k, b)$  are model parameters. These weights are further normalized to lie between 0 and 1. We recognize that this does not produce realistic probability estimates. The goal here is only to find a reasonable strategy to score and rank documents with respect to a boolean query.

## 4 Experimental Testing of the Model

We adopt the following experimental framework. Start with queries and documents in a single language. Convert the queries to another language using human translators and send both sets of queries to the CLIR system. From these experiments, one can determine the relative performance of cross-language IR compared to an ideal standard (monolingual retrieval), both overall and on a query by query basis. The per-query comparisons are indispensable for performing a failure analysis of the cross-language component of the system. Our experiments are based on the TREC El Norte collection, a set of Mexican newspaper articles written in Spanish. The collection comes with 50 topics and a large number of relevance judgements, although we only use the TREC-4 topics (numbers 26-50) in our experiments. We also take advantage of English translations of the topics generously provided by Davis and Dunning of NMSU in their TREC-4 paper [6]. The El Norte collection consists of about 200 MB of text, roughly 58,000 documents.

Our experiments use an extensively modified version of the SMART information retrieval system [3]. It is been updated to index 8-bit characters and the stemming algorithms have been replaced by the Xerox Spanish morphological analysis tools, which allow us to analyze the text and replace each term with its inflectional root. The SMART indexing functions are used basically unchanged, but the retrieval functions have been completely rewritten in order to implement the weighted boolean model. We also generate an extensive list of phrases which consist of adjacent non-stopword pairs, at least one of which is a noun (as determined by our Spanish part of speech tagger). Both the inflectional stemming and phrase generation are done as a preprocessing step prior to applying SMART to index the documents.

We have designed the following experiment to test the performance of the weighted boolean model for cross-language information retrieval. Boolean structured queries are constructed for the Spanish and English versions of the topics (in parallel by the author). The English queries are translated into Spanish using an on-line English-Spanish dictionary from Oxford University Press that has been specially modified for the text retrieval problem. Essentially, this involves the removal of all extraneous material from the dictionary entries except for

<b>Spanish:</b>	La fabricación en México de joyas de plata y oro.
<b>English:</b>	Silver and gold jewelry manufacturing in Mexico.
<b>Spanish Boolean:</b>	
	2 México mexicano
	2 joya joyas
	1 fabricación fabricar manufacturar
	3 oro plata
<b>English Boolean:</b>	
	2 Mexico Mexican
	2 jewelry
	1 manufacture manufacturing
	3 gold silver
<b>Retranslated Spanish Boolean:</b>	
	2 méxico mexicano méjico mejicano
	2 joya joyas alhaja
	1 fabricación manufactura fabricar manufacturar manufacturero fabril
	3 oro plata cubertería platear azogar blanquear

Table 1: Natural language and structured versions of query Q50.

the direct translations of each term. Unfortunately, the dictionary entries still contained some noise (and missing terms) due to the imperfect nature of the automatic cleaning process. We were concerned that interaction between dictionary errors and the models might seriously influence the experimental results, so we decided to clean up the translated queries. This was done manually by the author (not a Spanish speaker) with help from the 3rd edition Collins English-Spanish Dictionary. Unrelated terms were discarded and missing terms added, but no disambiguation was performed during the manual cleaning step. This means that the experiments presented here reflect ideal operating conditions which cannot currently be duplicated in our working CLIR system.

#### 4.1 A Sample Query

One sample topic is presented in all its forms in Table 1. We quickly note that the word “silver” is translated by both its noun and its verb forms. While we would like to perform part-of-speech disambiguation using our Spanish POS-tagger (as suggested by Davis [5]), it requires natural language text and cannot be applied to a structured query. This represents a key disadvantage of the weighted boolean model. One possible solution is to allow for manual disambiguation during the query construction process, although that option is not explored in our experiments. The English queries took at most a few minutes each to construct, while the Spanish queries took somewhat longer (due to the author’s lack of Spanish expertise). This means that there may well be errors

in the Spanish query formulations. The author used a bilingual dictionary to help with Spanish query construction (the same one used to manually clean the automatic translations), which could well be a more significant methodological problem. This may mean that the original and translated Spanish queries have more words in common than should happen in practice. However, this should not bias the results in favor of either of the two retrieval models.

## 4.2 Experimental Design, Results, and Analysis

In this experiment, we wish to test the claim that the weighted boolean model should perform proportionally better than the vector space model for cross-language information retrieval. It is important to limit the influence of external confounding factors, which suggests the following experimental design. Conduct four experimental runs as follows:

- (1) Weighted Boolean model on original Spanish queries
- (2) Weighted Boolean model on translated Spanish queries
- (3) Vector Space model on original Spanish queries
- (4) Vector Space model on translated Spanish queries

The key question is what to use for the vector space query. If we use the original natural language formulation, this gives a strong advantage to the boolean system, since its queries are manually constructed. Therefore, we decided instead to remove the structure from the boolean queries (e.g. treat them as a bag of words) for the runs using the vector model. However the concept weights are retained and applied to the vector model term weights in a direct multiplicative fashion. Therefore, runs (1)/(3) and (2)/(4) use exactly the same index terms and should be directly comparable based on the merits of the underlying models. Furthermore, the same term weighting scheme (BM25, as previously described) is used for both models.

One can view the results of these experiments as a 25x2x2 3-way table (query by model by translation). There are three effects of interest which can be measured from this design: model effect, translation effect, and the model-translation interaction. All of these effects are important and interesting to measure, but it is the significance of the model-translation interaction (i.e. testing whether (1)-(2)  $\ll$  (3)-(4)) which is necessary to verify our hypothesis. Note that we choose to model the queries as *random effects*, meaning that they are a random sample from an underlying population over which we have no control and therefore should only be considered as a random source of variation. This decision makes a difference for the statistical analysis of the results.

Finally, we must select an evaluation measure which is appropriate to the task. As described in our previous work [13], we believe that a high precision measure should be used for the cross-language adhoc search task. In general,

users of the system should not be expected to read, evaluate, or translate large numbers of documents in a foreign language. Another argument for high precision is that monolingual relevance feedback is likely to be much more effective than the original search formulation once one or more relevant documents have been found. Therefore, we choose to use precision averaged at 5, 10, 15, and 20 documents retrieved as the primary evaluation measure. However, in order to make our results roughly comparable to other work on the same collection, we also use uninterpolated average precision at the top 1000 documents, the traditional TREC evaluation measure.

	Boolean	Vector	average
monolingual	0.304/0.568	0.272/0.514	0.288/0.541
cross-language	0.281/0.541	0.202/0.415	0.242/0.478
average	0.292/0.554	0.237/0.464	0.265/0.509
boolean:	0.554-0.464		= 0.090
cross-language:	0.478-0.541		= -0.063
interaction:	(0.541-0.415)-(0.568-0.514)		= 0.072

Table 2: Average/high precision 2 by 2 score table averaged over 25 queries.

The average performance figures are presented in table 2. The results seem very promising as both the boolean effect and the interaction term are relatively large and positive, indicating that the boolean structured queries produce better performance, especially for the cross language problem. The effect of searching across languages is relatively small compared to what one might expect. There are two factors which probably account for this result: manual cleaning of the dictionary and the fact that the author used the same dictionary for cleaning and to help him generate the Spanish language queries. However, these results need to be validated with statistical testing. We perform an Analysis of Variance (ANOVA) on the 3-way table using a mixed model with two fixed effects (wtb = weighted boolean, clir = cross-language) and one random effect (qry = query), as shown in Table 3.

For details on the analysis of mixed models, the reader is referred to Lindman [16]. Unfortunately, the only effect which looks likely to be significant is the boost in performance provided by the weighted boolean model (p-value: 0.003). From the ANOVA results, we can conclude that our new weighted boolean model works better than the vector space model in general, but there is not enough evidence to say that it is particularly helpful for cross-language information retrieval. As these results are based on a relatively small sample of 25 queries, it seems reasonable to expect that a larger query sample may verify our hypothesis, and this represents our most immediate direction of future research. The performance of the weighted boolean model in general is quite impressive, although the short initial topic statements and the lack of query expansion bias the results somewhat in that direction. We also plan to explore our weighted

Factor	df	SS	MS(num)	MS(den)	F-val	Pr(F)
wtb	1	0.202	0.202	0.019	10.77	0.003
clir	1	0.098	0.098	0.033	3.00	0.096
qry	24	5.172	0.215			
wtb:clir	1	0.033	0.033	0.011	2.89	0.102
wtb:qry	24	0.450	0.019			
clir:qry	24	0.785	0.033			
wtb:clir:qry	24	0.271	0.011			

Table 3: ANOVA table: df = degrees of freedom, SS = sum of squares, MS = mean square (numerator/denominator), F-val = F statistic, Pr(F) = p-value.

boolean model further in a monolingual setting.

**Acknowledgments** The author would like to thank: **Kalervo Järvelin** for an interesting discussion which strongly motivated him to pursue this line of research, **Marian Ewell** for help with the statistical analysis of mixed models, **Maximilian Schulze** for support in document indexing and retrieval, and **Gregory Grefenstette** for help with dictionary construction and useful discussions on the CLIR problem.

## References

- [1] Lisa Ballesteros and W. Bruce Croft. Dictionary-based methods for cross-lingual information retrieval. In *Proc. of the 7th International DEXA Conference on Database and Expert Systems Applications*, 1996. To appear.
- [2] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- [3] Chris Buckley. Implementation of the smart information retrieval system. Technical Report 85-686, Cornell University, 1985. SMART is available for research use via anonymous FTP to ftp.cs.cornell.edu in the directory /pub/smart.
- [4] C.L.A. Clarke, G.V. Cormack, and F.J. Burkowski. Shortest substring ranking (multitext experiments for TREC-4). In *The 4th Text Retrieval Conference (TREC-4), NIST SP 500-236*, pages 295–304, 1996.
- [5] Mark Davis. New experiments in cross-language text retrieval at NMSU’s computing research lab. In *The 5th Text Retrieval Conference (TREC-5)*, 1997. To appear.

- [6] Mark Davis and Ted Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In *The 4th Text Retrieval Conference (TREC-4)*, NIST SP 500-236, pages 483–497, 1996.
- [7] Shinichi Doi and Kazunori Muraki. Translation ambiguity resolution based on text corpora of source and target languages. In *Proc. of the 15th Conference on Computational Linguistics (COLING '92)*, volume 2, pages 525–531, 1992.
- [8] E. Fox, S. Betrabet, M. Koushik, and W. Lee. *Information Retrieval - Data Structures and Algorithms*, chapter 15, pages 393–418. Prentice Hall, 1992.
- [9] Gregory Grefenstette, editor. *Workshop on Cross-Linguistic Information Retrieval - SIGIR '96*, 1996.
- [10] Donna Harman, editor. *The 4th Text Retrieval Conference (TREC-4)*, NIST SP 500-236, 1996.
- [11] Marti A. Hearst. Improving full-text precision on short queries using simple constraints. In *Symposium on Document Analysis and Information Retrieval (SDAIR)*. University of Nevada, Las Vegas, 1996.
- [12] David A. Hull. A probabilistic model for the approximate matching of boolean constraints. Being cleared for publication, 1997.
- [13] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proc. of the 19th ACM/SIGIR Conference*, pages 49–57, 1996.
- [14] Thomas K. Landauer and Michael L. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proc. of the 6th Conference of UW Centre for the New OED and Text Research*, pages 31–38, 1990.
- [15] Joon Ho Lee. Properties of extended boolean models in information retrieval. In *Proc. of the 17th ACM/SIGIR Conference*, pages 182–190, 1994.
- [16] Harold R. Lindman. *Analysis of Variances in Complex Experimental Designs*. Freeman and Co., 1974.
- [17] Douglas W. Oard. Alignment of spanish and english TREC topic descriptions. In *The 5th Text Retrieval Conference (TREC-5)*, 1997. To appear.
- [18] Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-9619, University of Maryland, 1996. <http://www.ee.umd.edu/medlab/mlir/mlir.html>.

- [19] D.W. Oard, N. DeClaris, B.J. Dorr, and C. Faloutsos. On automatic filtering of multilingual texts. In *Proc. of the 1994 IEEE Conference on Systems, Man, and Cybernetics*, 1994.
- [20] Carol Peters and Eugenio Picchi. Capturing the comparable: a system for querying comparable text corpora. In *Proc. of Analisi Statistica dei Dati Testuali (JADT)*, pages 247–254, 1995.
- [21] B.R. Pevzner. Comparative evaluation of the operation of the Russian and English variants of the Pusto-Nepusto-2 system. *Automatic Documentation and Mathematical Linguistics*, 6(2):71–74, 1972.
- [22] A. Steven Pollitt and Geoff Ellis. Multilingual access to document databases. In *Proc. of the 21st Conference of the Canadian Association for Information Science*, pages 128–140, 1993.
- [23] Khaled Radwan. *Vers l'Accès Multilingue en Langage Naturel aux Bases de Données Textuelles*. PhD thesis, Université de Paris-Sud, Centre d'Orsay, 1994.
- [24] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Overview of the 3rd Text Retrieval Conference (TREC-3), NIST SP 500-225*, 1995.
- [25] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21:187–194, 1970.
- [26] Páraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proc. of the 19th ACM/SIGIR Conference*, pages 58–65, 1996.
- [27] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.