

Monolingual, Bilingual, and GIRT Information Retrieval at CLEF-2005

Jacques Savoy, Pierre-Yves Berger

Institut interfacultaire d'informatique, Université de Neuchâtel,
Pierre-à-Mazel 7, 2001 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch Pierre-Yves.Berger@unine.ch

Abstract. For our fifth participation in the CLEF evaluation campaigns, our first objective was to propose an effective and general stopword list as well as a light stemming procedure for the Hungarian, Bulgarian and Portuguese (Brazilian) languages. Our second objective was to obtain a better picture of the relative merit of various search engines when processing documents in those languages. To do so we evaluated our scheme using two probabilistic models and five vector-processing approaches. In the bilingual track, we evaluated both the machine translation and bilingual dictionary approaches applied to automatically translate a query submitted in English into various target languages. Finally, using the GIRT corpora (available in English, German and Russian), we investigated the variations in retrieval effectiveness that resulted when we included or excluded manually assigned keywords attached to the bibliographic records (mainly comprising a title and an abstract).

1 Introduction

Since 2001, our research group has been investigating effective information retrieval (IR) techniques capable of handling a variety of natural languages [1], [2], in order to improve both monolingual and bilingual searches. Along this same stream, and based on our assumption that nouns and adjectives reveal the most about semantic content of documents (or requests), we designed a set of stopword lists and light stemming procedures for certain European languages. We then designed linguistic tools that would automatically remove inflectional suffixes attached to nouns and adjectives used to denote the gender (masculine, feminine, neutral), the number (singular or plural) and the case (nominative, dative, ablative, etc.). Needless to say we have also investigated other linguistic phenomena, such as compound constructions.

The rest of this paper is organized as follows: Section 2 outlines the main aspects of our stopword lists and light stemming procedures. Section 3 explains the principal features of different indexing and search strategies, and then evaluates them using the available corpora. The data fusion approaches used in our experiments and our official results are exposed in Section 4. Our bilingual experiments are presented and evaluated in Section 5, and Section 6 describes our experiments involving the domain-specific GIRT corpus.

2 Stopword Lists and Stemming Procedures

In order to define general stopwords lists, we first created a list of the top 200 most frequently occurring words, and then removed some words from this list (e.g., police, minister, president, Magyar). To this list we then added articles, pronouns, prepositions, conjunctions or very frequently occurring verb forms (e.g., to be, is, has, etc.). As a result of this procedure, we compiled a new stopword list for the Bulgarian and Hungarian languages (available at www.unine.ch/info/clef/). Thus, our final stopword list contained 463 words for the French language, 737 (761 in a previous version) for Hungarian, 258 (418 in a previous version) for Bulgarian, and 400 for Portuguese-Brazilian (we added eight Brazilian words to our Portuguese stopword list. These eight words are usually variants with or without accents, such as "vezes" in Portuguese and "vêzes" in Brazilian).

Once high-frequency words had been removed, our indexing procedure generally applied a stemming algorithm in order to conflate word variants into the same stem or root. Our first step in developing this procedure was to remove inflectional suffixes. For the Bulgarian language, we encountered some additional morphological difficulties. In this language, the definite article is usually represented by a suffix; for example, "mope" (sea) becomes "mopeto" (the sea) while "mopeta" (seas) becomes "mopetata" (the seas). For nouns, the general pattern is as follows: <stem><plural><article>. Contrary to other Slavic languages (such as Russian), Bulgarian does not add a suffix to indicate grammatical cases.

The Hungarian language shares certain similarities with the Finnish language (although these languages do not strictly belong to the same family, they can be viewed as cousins). Like Finnish, Hungarian has several number cases (usually 18) and each case has its own unambiguous form. For example, the noun "house" ("ház") may appear as "házat" (accusative case, as in "(I see) the house"), "házakat" (accusative plural case, as in "(I see) the houses"), "házamat" ("... my house") or "házaimat" ("... my houses"). In this language, the general construction used for nouns is as follows: <stem><plural><possessive marker><case>. For example, for <ház>a <m>a <t>in which the letter "a" is introduced to facilitate better pronunciation ("házmt" would be difficult to pronounce). From the IR point of view, some of Hungarian's linguistic features are viewed as good news. For example, a gender distinction is not attached to each noun (like in English) and adjectives are invariable, as in "... a szép ház" ("a beautiful house") or "... a szép házamat" ("my beautiful house"). Our suggested stemming procedures for these languages can be found at www.unine.ch/info/clef/.

Diacritic characters are usually not present in English collections (with certain exceptions, such as "résumé" or "cliché"). For the Hungarian, and Portuguese languages, we replaced these characters with their corresponding non-accentuated letters, even though the removal of accents from the Hungarian language can lead to some semantic ambiguity (e.g., between "kor" ("age") and "kór" ("illness"), or "ver" ("hurt") and "vér" ("blood")).

Finally, most European languages manifest other morphological characteristics, with compound word constructions being only one example (e.g., handgun, worldwide). In some experiments on Hungarian and German retrieval within the

GIRT corpus (Section 6), we used our own decomposing algorithm [3], leaving both the compound words and their component parts in the documents and queries.

3 Indexing and Searching Strategies

In order to obtain a broader view of the relative merit of various retrieval models, we first adopted the classical $tf \cdot idf$ weighting scheme (with cosine normalization), retrieval model denoted "doc=ntc, query=ntc" or "ntc-ntc"). To measure the similarity between documents and requests, we computed the inner product. Various other indexing weighting schemes have been suggested, as for example, the IR model denoted by "doc=Lnu" [4], "doc=dtu" [5].

In addition to these IR models based on the vector-space paradigm, we also considered probabilistic approaches such as the Okapi model [6]. As a second probabilistic approach, we implemented the Prosit model, one of a family of models suggested by Amati & Rijsbergen [7]. The exact specification of these IR models is given in [2].

To measure the retrieval performance, we adopted non-interpolated mean average precision (MAP). Then, to statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology [8]. Thus, in the tables in this paper we have underlined statistically significant differences resulting from the use of a two-sided non-parametric bootstrap test (significance level fixed at 5%).

We indexed the various collections using words as indexing units. The evaluations of our two probabilistic models and five vector-space schemes are listed in Table 1. In this table, the best performance under the given conditions is shown in bold type and it is used as a baseline for our statistical testing. The underlined results therefore indicate that the difference in mean average precision can be viewed as statistically significant when compared to the best system value. As can be seen in the top part of Table 1, the Okapi model was usually the best IR model for the French and Portuguese collections. For these two corpora however, the MAP differences between the various IR models are not always statistically significant. The Prosit model performs best result for the Bulgarian collection, while for the Hungarian corpus, the Okapi probabilistic approach was the solution that performed best (bottom part of Table 1). For this same language, statistics for five IR models revealed similar performance levels (Okapi, Prosit, "Lnu-ltc", " dtu-dtn", "atn-ntc"). However, overall statistics like the MAP may hide performance irregularities among queries, and in this regard Tomlinson [9] presented examples demonstrating that, while a given search strategy may improve retrieval performance for some queries, it may lead to decreases for others.

Moreover, the data in Table 1 shows that when the number of search terms increases (from T, TD to TDN), retrieval effectiveness usually increases also. The average improvement is of about 33.4% result when comparing title-only

Table 1. MAP of single searching strategies

Mean average precision						
Query Model	French T 50 queries	French TD 50 queries	French TDN 50 queries	Portug. T 50 queries	Portug. TD 50 queries	Portug. TDN 50 queries
Prosit	<u>0.2895</u>	0.3696	0.3961	<u>0.2755</u>	0.3438	0.3697
Okapi	0.3029	0.3754	0.3948	0.2873	0.3477	0.3719
Lnu-ltc	<u>0.2821</u>	<u>0.3437</u>	<u>0.3703</u>	<u>0.2611</u>	0.3338	<u>0.3517</u>
dtu-dtn	<u>0.2726</u>	<u>0.3365</u>	<u>0.3633</u>	<u>0.2571</u>	0.3221	<u>0.3338</u>
atn-ntc	<u>0.2809</u>	<u>0.3328</u>	<u>0.3507</u>	<u>0.2458</u>	<u>0.3076</u>	<u>0.3433</u>
ltn-ntc	<u>0.2588</u>	<u>0.3066</u>	<u>0.3232</u>	<u>0.2149</u>	<u>0.2535</u>	<u>0.2740</u>
ntc-ntc	<u>0.1862</u>	<u>0.2175</u>	<u>0.2335</u>	<u>0.1553</u>	<u>0.1868</u>	<u>0.2221</u>
Query Model	Bulgarian T 49 queries	Bulgarian TD 49 queries	Bulgarian TDN 49 queries	Hungarian TD 50 queries	Hungarian TD-decomp 50 queries	Hungarian TD-light 50 queries
Prosit	0.2662	0.3030	0.3132	0.3420	0.3390	0.3359
Okapi	<u>0.2350</u>	0.2760	0.2819	0.3501	0.3391	0.3410
Lnu-ltc	<u>0.2268</u>	0.2737	0.2800	0.3301	0.3273	0.3249
dtu-dtn	<u>0.2288</u>	<u>0.2575</u>	<u>0.2522</u>	0.3401	0.3341	0.3280
atn-ntc	<u>0.2340</u>	<u>0.2618</u>	<u>0.2578</u>	0.3215	0.3179	0.3199
ltn-ntc	<u>0.1679</u>	<u>0.2031</u>	<u>0.2076</u>	<u>0.2853</u>	<u>0.2820</u>	0.2856
ntc-ntc	<u>0.1781</u>	<u>0.1967</u>	<u>0.2074</u>	<u>0.2208</u>	<u>0.2099</u>	<u>0.2245</u>

(or T) with TDN queries for the Portuguese collection, 31.3% when comparing the French corpus, and 6.4% for the Bulgarian collection.

With the Hungarian collection, we automatically decomposed long words (composed by more than 8 characters) using our own algorithm [3]. In this experiment, both the compound words and their components were left in both documents and queries (under the label "TD-decomp" in the bottom part of Table 1). Using the TD queries and the Okapi model, we obtained a MAP of 0.3391, revealing a decrease of 3.1% when compared to an indexing approach that did not use decomposing (0.3501). Based on the five best retrieval schemes, the average performance decrease was around 1.6%. Using a lighter stemmer (fewer rules) for the Hungarian language (retrieval performance listed under the label "TD-light" in Table 1), the average difference in MAP over the five best retrieval schemes was around 2%, and in favor of the original stemming approach. Tordai & de Rijke [10] also evaluated various stemming algorithms for the Hungarian languages, finding that a light stemming approach might prove effective for a morphologically rich language such as Hungarian.

It has been observed that pseudo-relevance feedback (PRF or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [4] with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add m terms extracted from the k best ranked documents from the original query. To evaluate this proposition, we used the Okapi

and the Prosit probabilistic models and enlarged the query by the 10 to 20 terms retrieved from the 3 to 10 best-ranked articles.

Table 2 depicted the best results obtained with the PRF technique for the Okapi model. This demonstrates that the optimal parameter setting seemed to be collection-dependant. Moreover, performance improvement also seemed to be collection-dependant, with the French corpus showing an increase of 9.2% (from a mean average precision of 0.3754 to 0.4099), 5.2% for the Portuguese collection (from 0.3477 to 0.3668), 1.3% for the Hungarian collection (from 0.3501 to 0.3545), and 0.8% for the Bulgarian corpus (from 0.2704 to 0.2726). In Table 2, the baseline used for our statistical testing was the MAP calculated before the query was automatically expanded. In this case, it is interesting to note that our statistical testing does always detect any significant difference.

Table 2. MAP using blind-query expansion (Okapi model)

Query TD Model	Mean average precision			
	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
Okapi	0.3754	0.3477	0.2760	0.3501
k docs/ m terms	3/10 <u>0.3967</u> 5/15 <u>0.4034</u> 10/15 0.4099 10/20 <u>0.4075</u>	3/15 <u>0.3656</u> 5/15 0.3668 10/15 0.3626 10/20 0.3601	3/15 <u>0.2500</u> 5/15 0.2553 10/15 0.2778 10/20 0.2718	3/10 0.3545 5/10 0.3513 5/15 0.3490 10/15 0.3492

4 Data Fusion and Official Results

It is assumed that combining different search models should improve retrieval effectiveness, due to the fact that different document representations might retrieve different pertinent items and thus increase the overall recall [11]. On the other hand, when combining different search schemes, we might suppose that these various IR strategies are more likely to rank the same relevant items higher on the list than they would non-relevant documents (viewed as outliers). In this current study we combined the two probabilistic models Okapi and Prosit using the data fusion operators defined in [2].

Table 3 shows the exact specifications of our best-performing official monolingual runs. In these experiments, we combined the Okapi and the Prosit probabilistic models using the Z-Score (see [2]) data fusion operator for the French and Portuguese corpora. For the Hungarian and Bulgarian collection, our best results were achieved using the Prosit model (see Table 3).

Table 3. Description and MAP of our best official monolingual runs

Run name	Lan.	Query	Model	Query exp.	Combined	MAP
UniNEfr1	French	TD	Okapi	3 docs/10 terms	Z-scoreW	0.4207
		TD	Prosit	5 docs/50 terms		
UniNEpt2	Portug.	TD	Okapi	3 docs/15 terms	Z-scoreW	0.3875
		TD	Prosit	5 docs/60 terms		
UniNEbg3	Bulgarian	TD	Prosit	5 docs/30 terms	n/a	0.2839
UniNEhu3	Hungarian	TD	Prosit	5 docs/40 terms	n/a	0.3889

5 Bilingual Information Retrieval

For the bilingual track, we chose English as the language to be used for submitting queries for automatic translation into the four different languages. We used seven different machine translation (MT) systems and three bilingual dictionaries ("Babylon," "Ectaco," and "Medios"). The freely available translation tools used in our experiments are listed below:

1. SYSTRAN www.systranlinks.com
2. GOOGLE www.google.com/language_tools
3. FREETRANSLATION www.freetranslation.com
4. INTERTRAN www.tranexp.com/
5. WORLDLINGO www.worldlingo.com/
6. BABELFISH babelfish.altavista.com/
7. PROMT webtranslation.paralink.com/
8. BABYLON www.babylon.com
8. ECTACO www.ectaco.co.uk/free-online-dictionaries
10. MEDIOS consulting.medios.fi/dictionary.

When using the different bilingual dictionaries to translate an English request word-by-word, more than one translation was usually provided, in an unspecified order. We thus decided to pick either the first translation available (labeled "Babylon 1" or "Ectaco 1") or the first two terms available (labeled "Babylon 2").

Our experiments show that Google provided the best translation for the French collection and Prompt for the Portuguese corpus. The FreeTranslation and Prompt MT systems usually obtain satisfactory retrieval performances for both these languages. For French, the BabelFish and Systran translation systems worked well. For Bulgarian and Hungarian, we found only a few translation tools, and unfortunately their overall performance levels were not very good.

Table 4 shows the retrieval effectiveness for various query translation combinations when using the Okapi probabilistic model. The top part of the table indicates the exact query translation combination used while the bottom part shows the MAP obtained with our combined query translation approach. In order to select which query translations would be combined, we made use of our prior findings [2] as well as our own intuition before selecting best translation

tools. As can be seen in Table 4, the resulting retrieval performances depicted are sometimes better than the best single translation scheme, as shown in the row labeled "Best single" (e.g., the "Comb 1" strategy for French, or the "Comb 3" or "Comb 5" strategies for Portuguese, "Comb 2" for Bulgarian, and "Comb 5" for Hungarian). From a statistical perspective however these combined query translation approaches did not perform better than the best single translation tool (except "Comb 3" for the Portuguese corpus).

Table 4. MAP of various combined translation devices (Okapi model)

TD queries Model	Mean average precision			
	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
Comb 1	Systran+Promt	Promt+Bab 1	Inter+all 2	Inter+Ecta 1
Comb 2	Lingo+Bab 1	Promt+Inter	Ecta 1+Bab 2	Inter+Bab 1
Comb 3	Free+Promt +Babylon 1	Prompt+Free +Babylon 1		Bab 1+Med 2 +Ectaco 1
Comb 4	Lingo+Promt +Babylon 1	Prompt+Inter +Babylon 1	Inter+Ecta 1 +Babylon 2	Inter+Bab 1 +Ectaco 1
Comb 5		Prompt+Free +Inter+Bab 1		Inter+Bab 1+ Med 2+Ecta 1
Best single	0.3259	0.2673	0.0800	0.1822
Comb 1	0.3274	0.2849	0.0831	0.1845
Comb 2	0.3089	0.2749	0.0962	0.1876
Comb 3	0.3246	<u>0.2977</u>		0.1966
Comb 4	0.3228	0.2955	0.0908	0.2005
Comb 5		0.2978		0.2183

Finally, Table 5 lists the parameter settings used for our best performing official runs in the bilingual task. For each experiment, queries were written in English in order to retrieve documents in the other target languages. Before combining the result lists we automatically expanded the translated queries using a pseudo-relevance feedback method (Rocchio's approach in this case).

6 Monolingual Domain-Specific Retrieval: GIRT

In the domain-specific retrieval task (called GIRT), the three available corpora are composed of bibliographic records extracted from various sources in the social sciences domain. These collections contain a total of 397,218 documents or about 590 MB, written for the most part in German. A typical record in this collection contains a title, an abstract, a set of manually assigned keyword, and some additional information of less importance from an IR perspective (e.g., authors' name, publication date, etc.). The GIRT corpus thus allowed us to

Table 5. Description and MAP of our best official bilingual runs

From EN to ...	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
IR 1 (<i>k d./m t.</i>)	Okapi (10/10)	Okapi (10/30)	Prosit (3/50)	Prosit (3/50)
IR 2 (<i>k d./m t.</i>)		Prosit (10/20)		
Data fusion		Z-scoreW		
Translation tools	Comb3	Comb4	Comb3	Comb5
MAP	0.3467	0.3404	0.1399	0.2882
Run name	UniNEbifr2	UniNEbipt1	UniNEbibg3	UniNEbihu3

evaluate the impact of manually assigned descriptors and compare them to an indexing scheme, based only on the information contained in the corresponding article’s title and abstract sections. To tackle this we evaluated all of the GIRT collection (denoted "all" in Table 6) or only the titles and abstracts taken from the bibliographic records (under the label "TI & AB"). In our experiments, the decrease in mean average precision was around 14.4% for the German corpus and 36.5% for the English GIRT collection.

Table 6. MAP of various single searching strategies (GIRT corpus)

Language Query TD Model	Mean average precision				
	German all 25 queries	German TI & AB 25 queries	English all 25 queries	English TI & AB 25 queries	Russian all 25 queries
Prosit	<u>0.4249</u>	0.3659	0.4645	0.2948	<u>0.2270</u>
Okapi	0.4353	0.3645	0.4604	0.2854	0.2742
Lnu-ltc	<u>0.3977</u>	<u>0.3307</u>	0.4234	<u>0.2712</u>	0.2577
dtu-dtn	<u>0.3789</u>	<u>0.3236</u>	<u>0.3936</u>	0.2738	0.3003
atn-ntc	<u>0.3914</u>	<u>0.3458</u>	<u>0.4102</u>	<u>0.2681</u>	0.2695
ltn-ntc	<u>0.3724</u>	<u>0.3146</u>	<u>0.3448</u>	<u>0.2158</u>	<u>0.2636</u>
ntc-ntc	<u>0.2765</u>	<u>0.2452</u>	<u>0.2859</u>	<u>0.2023</u>	<u>0.1393</u>

Our best performing official runs in the monolingual GIRT task are listed in Table 7. For each language, we submitted the first run using a data fusion operator ("Z-ScoreW" in this case). For all runs, we automatically expanded the queries using a blind relevance feedback method (Rocchio’s in our experiments), hopping to improve retrieval effectiveness.

Table 7. Description and MAP of our best official GIRT runs

Run name	Lan.	Query	Model	Query exp.	Combined	MAP
UniNEgde1	GE	TD	Okapi	5 d. / 10 t.	Z-scoreW	0.4921
		TD	Prosit	10 d. / 125 t.		
UniNEgen1	EN	TD	Okapi	5 d. / 10 t.	Z-scoreW	0.5065
		TD	Prosit	10 d. / 50 t.		
UniNEgru2	RU	TD	Okapi	5 d. / 20 t.	n/a	0.2774

7 Conclusion

In this sixth CLEF evaluation campaign, we proposed a general stopword list and a light stemming procedure (removing only inflections attached to nouns and adjectives) for the Bulgarian and Hungarian languages. Based on two different probabilistic IR models and five vector-processing schemes (see Table 1), we found that the Okapi or the Prosit models provide the best retrieval performances for all the different languages. Compared to the classical $tf \cdot idf$ model, this approach results in mean average precision improvements of 72% for the French corpus (TD queries, Okapi), 86% for the Portuguese (TD queries, Okapi), 58% for the Hungarian (TD queries, Okapi), and 54% for the Bulgarian (TD queries, Prosit). When query size is increased from title-only (T) to the longest request formulation (TDN), retrieval performance is also increased (33% for Portuguese, 31% for French, 21% for Hungarian).

As in previous evaluation campaigns we were able to confirm that pseudo-relevance feedback based on Rocchio’s model would usually improve mean average precision for the French and Portuguese language, even though this improvement is not always statistically significant (see Table 2). For the other languages (Bulgarian and Hungarian), this blind query expansion did not improve mean average precision from a statistical point of view. In an effort to hopefully enhance retrieval performance, we could use a data fusion approach to combine two or more IR models. The use of this search strategy did however require building two inverted files, thus doubling the search time needed.

The automatic decomposing of Hungarian words and its impact in IR remains an open question and our preliminary experiments provide no clear and precise answers (our decomposing scheme did however decrease retrieval performance slightly, as shown in bottom part of Table 1).

In the bilingual task, the freely available translation tools perform reasonably well for both the French and Portuguese languages (based on the three best translation tools, the MAP compared to the monolingual search is around 85% for the French language and 72.6% for the Portuguese). For the less frequently used languages Bulgarian and Hungarian, the freely available translation tools (either the bilingual dictionary or the MT system) do not perform well. Their MAP is around 50% for Hungarian, and 30% for Bulgarian compared to the retrieval performance of a monolingual search.

In the GIRT task (Table 6), the probabilistic models (either Okapi or Prosit) usually results in better retrieval performances. Moreover, when taking manually assigned descriptors into account, mean average precision improves by around 36.5% for the English corpus and 14.4% for the German collection.

Acknowledgments. The authors would like to also thank the CLEF-2005 task organizers for their efforts in developing various European language test-collections, and C. Buckley from SabIR for giving us the opportunity to use the SMART system. This research was supported in part by the Swiss National Science Foundation under Grant #21-66 742.01.

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. *IR Journal*, **7** (2004) 121–148
2. Savoy, J.: Data Fusion for Effective European Monolingual Information Retrieval. In: Peters, C., Clough, P.D., Jones, G.J.F., Gonzalo, J., Kluck, M., Magnini, B. (Eds.): *Multilingual Information Access for Text, Speech and Images. Lecture Notes in Computer Science: Vol. 3491.* Springer, Heidelberg (2005), 233–244
3. Savoy, J.: Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.): *Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science: Vol. 3237.* Springer, Heidelberg (2004), 322–336
4. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In *Proceedings TREC-4. NIST Publication #500-236*, Gaithersburg (1996) 25–48
5. Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F.: AT&T at TREC-7. In *Proceedings TREC-7. NIST, Publication #500-242*, Gaithersburg (1999) 239–251
6. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management*, **36** (2000) 95–108
7. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems*, **20** (2002) 357–389
8. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management*, **33** (1997) 495–512
9. Tomlinson, S.: European Ad Hoc Retrieval Experiments with Hummingbird SearchServerTM at CLEF 2005. In *this volume*
10. Tordai, A., de Rijke, M.: Hungarian Monolingual Retrieval at CLEF 2005. In *this volume*
11. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. *IR Journal*, **1** (1999) 151–173