German, French, English and Persian Retrieval Experiments at CLEF 2008

Stephen Tomlinson
Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com
http://www.opentext.com/

August 18, 2008

Abstract

We describe evaluation experiments conducted by submitting retrieval runs for the monolingual German, French, English and Persian (Farsi) information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2008. In the ad hoc retrieval tasks, the system was given 50 natural language queries, and the goal was to find all of the relevant records or documents (with high precision) in a particular document set. We conducted diagnostic experiments with different techniques for matching word variations, comparing the performance on the robust Generalized Success@10 measure and the non-robust mean average precision measure. The measures generally agreed on the mean benefits of morphological techniques such as decompounding and stemming, but generally disagreed on the blind feedback technique, though not all of the mean differences were statistically significant. Also, for each language, we submitted a sample of the first 10000 retrieved items to investigate the frequency of relevant items at deeper ranks than the official judging depth of 60. The results suggest that, on average, the percentage of relevant items assessed was less than 55% for each of German, French and English and less than 25% for Persian.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

German Retrieval, French Retrieval, English Retrieval, Persian Retrieval, Robust Retrieval, Sampling

1 Introduction

Livelink ECM - eDOCS SearchServer™ is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other components of the Livelink ECM - eDOCS Suite[1].

[1]Livelink, Open Text™ and SearchServer™ are trademarks or registered trademarks of Open Text Corporation in the United States of America, Canada, the European Union and/or other countries. This list of trademarks is not exhaustive. Other trademarks, registered trademarks, product names, company names, brands and service names mentioned herein are property of Open Text Corporation or other respective owners.

Table 1: Sizes of CLEF 2008 Ad-Hoc Track Test Collections

Code	Language	Text Size (uncompressed)	Documents	Topics	Rel/Topic
DE	German	1,306,492,248 bytes	869,353	50	33 (lo 2, hi 84)
EN	English	1,208,383,351 bytes	1,000,100	50	51 (lo 7, hi 190)
FA	Persian	628,471,252 bytes	166,774	50	103 (lo 7, hi 255)
FR	French	1,362,122,091 bytes	1,000,100	50	27 (lo 3, hi 224)

SearchServer works in Unicode internally [4] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (CLEF [3], NTCIR [5] and TREC [9]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This paper describes experimental work with SearchServer for the task of finding relevant documents for natural language queries in various European languages using the CLEF 2008 Ad-Hoc Track test collections.

2 Methodology

2.1 Data

The CLEF 2008 Ad-Hoc Track document sets consisted of tagged (XML-formatted) records or documents in 4 different languages: German, French, English and Persian (also known as Farsi). For German, French and English, the records were library catalog cards (bibliographic records describing publications archived by The European Library (TEL)). For Persian, the documents were newspaper articles (Hamshahri corpus of 1996-2002). Table 1 gives the collection sizes.

The CLEF organizers created 50 natural language "topics" (numbered 451-500 for German, French and English and 551-600 for Persian) and translated them into many languages. Sometimes topics are discarded for some languages because of a lack of relevant documents (though that did not happen this year). Table 1 gives the final number of topics for each language and their average number of relevant documents (along with the lowest and highest number of relevant documents of any topic). For more information on the CLEF test collections, see the track overview paper.

2.2 Indexing

Our indexing approach was mostly the same as last year [16]. Accents were not indexed. The apostrophe was treated as a word separator (except in English). The custom text reader, cTREC, was updated to maintain support for the CLEF guidelines of only indexing specifically tagged fields.

For some experiments, some stop words were excluded from indexing (e.g. words like "the", "by" and "of" in English). For our Persian experiments, our stop word list was based on Savoy's list [8].

By default, the SearchServer index supports both exact matching (after some Unicode-based normalizations, such as decompositions and conversion to upper-case) and morphological matching (e.g. inflections, derivations and compounds, depending on the linguistic component used).

2.3 Searching

We experimented with the SearchServer CONTAINS predicate. Our test application specified SearchSQL to perform a boolean-OR of the query words. For example, for German topic 451 whose Title was "Römisches Militär in Britannien" (Roman Military in Britain), a corresponding SearchSQL query would be:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM CLEF08DE
WHERE FT_TEXT CONTAINS 'Römisches'|'Militär'|'in'|'Britannien'
ORDER BY REL DESC;
```

Most aspects of the SearchServer relevance value calculation are the same as described last year [16]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [7] and dampens the inverse document frequency using an approximation of the logarithm. These calculations are based on the stems of the terms (roughly speaking) when doing morphological searching (i.e. when SET TERM_GENERATOR 'word!ftelp/inflect/decompound' was previously specified). The SearchServer RELEVANCE_METHOD setting was set to '2:3' and RELEVANCE_DLEN_IMP was set to 500 for all experiments in this paper.

2.4 Diagnostic Runs

For the diagnostic runs listed in Tables 2, the run names consist of a language code ("DE" for German, "EN" for English, "FA" for Persian, and "FR" for French) followed by one of the following labels:

- "none": No linguistic variations from stemming were matched. Just the surface forms were searched on (after case-normalization).
- "lexstem": Same as "none" except that linguistic variations from stemming were matched. The lexicon-based inflectional stemmer in SearchServer was used. For German, this stemmer includes decompounding.
- "algstem": Same as "lexstem" except that an algorithmic stemmer was used. For Persian, our stemmer was ported from Savoy's [8]. For German, French and English, Porter's algorithmic "Snowball" stemmers [6] were used (for English, the Porter2 version was used).
- "algall": Same as "algstem" except that a separate index was used which did not stop any words from being indexed.
- "4gram": Same as "lexall" except that the run used a different index which primarily consisted of the 4-grams of terms, e.g. the word 'search' would produce index terms of 'sear', 'earc' and 'arch'. No stemming was done; searching used the IS_ABOUT predicate (instead of the CONTAINS predicate) with morphological options disabled to search for the 4-grams of the query terms.

Note that all diagnostic runs just used the Title field of the topic.

2.5 Retrieval Measures

Traditionally, different retrieval measures have been used for "ad hoc" tasks, which seek relevant items for a topic, than for "known-item" tasks, which seek a particular known document. However, we argue that the known-item measures are not only applicable to ad hoc tasks, but that they are often preferable. For many ad hoc tasks, e.g. finding answer documents for questions, just one relevant item is needed. Also, the traditional ad hoc measures encourage retrieval of duplicate relevant documents, which does not correspond to user benefit.

The traditional known-item measures are very coarse, e.g. Success@10 is 1 or 0 for each topic, while reciprocal rank cannot produce a value between 1.0 and 0.5. In 2005, we began investigating a new measure, Generalized Success@10 (GS10) (introduced as "First Relevant Score" (FRS) in [13]), which is defined below. This investigation led to the discovery that the blind feedback technique (a commonly used technique at CLEF, NTCIR and TREC, but not known to be popular in real systems) had the downside of pushing down the first relevant item (on average), as has now

been verified not just for our own blind feedback approach, but for the 7 blind feedback systems of the 2003 RIA Workshop [11] and for the Neuchâtel system using French data from CLEF [1]. [2] provides a theoretical explanation for why positive feedback approaches are detrimental to the rank of the first relevant item.

2.5.1 Primary Recall Measures

"Primary recall" is retrieval of the first relevant item for a topic. Primary recall measures include the following:

- *Generalized Success@30* (GS30): For a topic, GS30 is 1.024^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found.
- *Generalized Success@10* (GS10): For a topic, GS10 is 1.08^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found.
- *Success@n* (S@n): For a topic, Success@n is 1 if a desired page is found in the first n rows, 0 otherwise. This paper lists Success@1 (S1) and Success@10 (S10) for all runs.
- *Reciprocal Rank* (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found. "Mean Reciprocal Rank" (MRR) is the mean of the reciprocal ranks over all the topics.

Interpretation of Generalized Success@n: GS30 and GS10 are estimates of the percentage of potential result list reading the system saved the user to get to the first relevant item, assuming that users are less and less likely to continue reading as they get deeper into the result list.

Comparison of GS10 and Reciprocal Rank: Both GS10 and RR are 1.0 if a desired page is found at rank 1. At rank 2, GS10 is just 7 points lower (0.93), whereas RR is 50 points lower (0.50). At rank 3, GS10 is another 7 points lower (0.86), whereas RR is 17 points lower (0.33). At rank 10, GS10 is 0.50, whereas RR is 0.10. GS10 is greater than RR for ranks 2 to 52 and lower for ranks 53 and beyond.

Connection of GS10 to Success@10: GS10 is considered a generalization of Success@10 because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$. (Similarly, GS30 is considered a generalization of Success@30 because it rounds to 1 for $r \leq 30$ and to 0 for $r > 30$.)

2.5.2 Secondary Recall Measures

"Secondary recall" is retrieval of the additional relevant items for a topic (after the first one). Secondary recall measures place most of their weight on these additional relevant items.

- *Precision@n*: For a topic, "precision" is the percentage of retrieved documents which are relevant. "Precision@n" is the precision after n documents have been retrieved. This paper lists Precision@10 (P10) for all runs.
- *Average Precision* (AP): For a topic, AP is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant documents which are not retrieved). By convention, AP is based on the first 1000 retrieved documents for the topic. The score ranges from 0.0 (no relevants found) to 1.0 (all relevants found at the top of the list). "Mean Average Precision" (MAP) is the mean of the average precision scores over all of the topics (i.e. all topics are weighted equally).
- *Geometric MAP* (GMAP): GMAP (introduced in [17]) is based on "Log Average Precision" which for a topic is the natural log of the max of 0.00001 and the average precision. GMAP is the exponential of the mean log average precision. (We argue in [11] that primary recall measures better reflect robustness than GMAP.)

Table 2: Mean Scores of Diagnostic Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
DE-lexstem	0.916	0.864	45/50	0.699	29/50	0.444	0.168	0.294
DE-4gram	0.909	0.831	43/50	0.658	27/50	0.394	0.129	0.249
DE-algstem	0.794	0.726	37/50	0.607	26/50	0.340	0.045	0.184
DE-none	0.756	0.675	34/50	0.547	23/50	0.252	0.022	0.124
EN-lexstem	0.936	0.879	45/50	0.757	33/50	0.492	0.156	0.283
EN-algstem	0.927	0.869	44/50	0.752	33/50	0.482	0.160	0.298
EN-none	0.897	0.834	42/50	0.732	33/50	0.444	0.122	0.248
EN-4gram	0.826	0.739	37/50	0.640	29/50	0.380	0.089	0.236
FA-4gram	0.979	0.946	48/50	0.852	39/50	0.642	0.358	0.425
FA-algall	0.977	0.942	47/50	0.859	40/50	0.640	0.356	0.425
FA-none	0.976	0.938	48/50	0.844	39/50	0.636	0.351	0.419
FA-algstem	0.974	0.934	48/50	0.841	39/50	0.626	0.350	0.421
FR-algstem	0.875	0.768	38/50	0.591	23/50	0.320	0.118	0.246
FR-lexstem	0.864	0.751	39/50	0.558	22/50	0.312	0.095	0.237
FR-none	0.853	0.742	39/50	0.562	23/50	0.294	0.080	0.219
FR-4gram	0.834	0.735	39/50	0.599	26/50	0.294	0.096	0.236

2.6 Statistical Significance Tables

For tables comparing 2 diagnostic runs (such as Table 3), the columns are as follows:

- "Expt" specifies the experiment. The language code is given, followed by the labels of the 2 runs being compared. (We abbreviate "lexstem" to "lex", "algstem" to "alg", "4gram" to "4gr" and "algall" to "all".) The difference is the first run minus the second run. For example, "DE-lex-none" specifies the difference of subtracting the scores of the German 'none' run from the German 'lexstem' run (of Table 2).
- "ΔGS10" is the difference of the mean GS10 scores of the two runs being compared (and "ΔMAP" is the difference of the mean average precision scores).
- "95% Conf" is an approximate 95% confidence interval for the difference (calculated from plus/minus twice the standard error of the mean difference). If zero is not in the interval, the result is "statistically significant" (at the 5% level), i.e. the feature is unlikely to be of neutral impact (on average), though if the average difference is small (e.g. <0.020) it may still be too minor to be considered "significant" in the magnitude sense.
- "vs." is the number of topics on which the first run scored higher, lower and tied (respectively) compared to the second run. These numbers should always add to the number of topics.
- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets. The first difference is the largest one of any topic (based on the absolute value). The third difference is the largest difference in the other direction (so the first and third differences give the *range* of differences observed in this experiment). The middle difference is the largest of the remaining differences (based on the absolute value).

Table 3: Impact of Stemming on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
DE-lex-none	0.189	(0.077, 0.300)	17-9-24	1.00 (488), 1.00 (455), −0.42 (468)
EN-lex-none	0.045	(−0.013, 0.102)	6-5-39	0.86 (491), 0.85 (453), −0.39 (487)
FR-lex-none	0.009	(−0.041, 0.059)	12-12-26	0.91 (461), 0.25 (466), −0.54 (458)
FA-alg-none	−0.004	(−0.013, 0.004)	2-5-43	−0.13 (599), −0.07 (557), 0.07 (585)
	ΔMAP			
DE-lex-none	0.170	(0.111, 0.230)	47-2-1	0.85 (460), 0.74 (493), −0.13 (456)
EN-lex-none	0.035	(0.001, 0.069)	27-16-7	0.63 (479), 0.36 (480), −0.15 (463)
FR-lex-none	0.018	(−0.009, 0.046)	25-23-2	0.48 (479), 0.20 (470), −0.16 (486)
FA-alg-none	0.002	(−0.009, 0.012)	17-30-3	0.13 (565), 0.09 (590), −0.08 (591)

3 Results of Morphological Experiments

3.1 Impact of Stemming

Table 3 shows the impact of stemming for the 4 languages. The mean increase in GenS@10 was statistically significant for German, and the mean increases in MAP were statistically significant for German and English.

Table 3 also shows that there were large impacts from stemming on particular topics for German, French and English in both the GenS@10 and MAP measures (we look at some examples in the later sections).

Surprisingly, for Persian, even on individual topics there was relatively little impact from stemming. We notice in Table 2 that the Success@10 rate was relatively high for Persian (48 out of 50) even without stemming, and that relevant documents were plentiful (103 per topic on average as per Table 1), but we have not done sufficient analysis to understand why the stemming impact was so minor across topics.

3.2 Lexical vs. Algorithmic Stemming

Table 4 isolates the differences between the lexical and algorithmic stemmers for the 3 languages for which both types of stemmers were available. For each language, each stemmer substantially outscored the other on at least some individual topics. The higher mean scores of lexical stemming for German were statistically significant in both the GenS@10 and MAP measures.

German is a language with frequent compound words, and the lexical stemmer included decompounding, unlike the algorithmic stemmer. For example, in German topic 455 (Irische Emigration nach Nordamerika (Irish Emigration to North America)) the run using lexical stemming returned a record mentioning "Massenemigration nach Nordamerika" first, in part because it produced 'emigration' as a stem of 'Massenemigration'. The run using algorithimic stemming did not recognize the common stem of Emigration and Massenemigration and did not return a relevant document until rank 463 (hence the huge difference in the GenS@10 score for this topic listed in Table 4).

Unfortunately, we haven't had time to walk through more of the stemming differences, but in the past we found a lot of them were from the lexical stemmers just matching inflections while the algorithmic stemmers often additionally match derivations [14].

3.3 Comparison to 4-grams

Table 5 compares the 4-gram results to stemming results for all 4 languages. While most of the mean differences are not statistically significant, there are a lot of large differences on individual topics.

Table 4: Lexical vs. Algorithmic Stemming in GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
DE-lex-alg	0.138	(0.031, 0.245)	12-10-28	1.00 (488), 1.00 (455), −0.54 (468)
EN-lex-alg	0.010	(−0.006, 0.025)	3-1-46	0.36 (456), 0.07 (455), −0.00 (473)
FR-lex-alg	−0.017	(−0.066, 0.032)	8-4-38	−1.00 (482), −0.43 (492), 0.33 (484)
	ΔMAP			
DE-lex-alg	0.111	(0.053, 0.168)	39-10-1	0.74 (493), 0.69 (490), −0.29 (458)
EN-lex-alg	−0.015	(−0.043, 0.012)	28-10-12	−0.64 (482), −0.19 (485), 0.03 (463)
FR-lex-alg	−0.009	(−0.028, 0.011)	25-12-13	−0.32 (492), −0.31 (478), 0.08 (496)

Table 5: Stems vs. 4-grams in GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
EN-lex-4gr	0.139	(0.053, 0.225)	17-7-26	0.99 (451), 0.97 (495), −0.19 (478)
DE-lex-4gr	0.033	(−0.035, 0.101)	14-12-24	−0.86 (497), 0.66 (451), 0.79 (471)
FR-lex-4gr	0.016	(−0.064, 0.096)	16-14-20	−1.00 (482), 0.73 (473), 0.92 (481)
FA-alg-4gr	−0.012	(−0.034, 0.010)	3-6-41	−0.50 (559), −0.08 (599), 0.07 (585)
FA-alg-all	−0.009	(−0.029, 0.012)	1-1-48	−0.50 (559), 0.00 (553), 0.07 (599)
	ΔMAP			
EN-lex-4gr	0.047	(−0.009, 0.102)	33-17-0	−0.68 (482), 0.40 (481), 0.58 (479)
DE-lex-4gr	0.045	(−0.003, 0.094)	31-19-0	0.61 (484), 0.53 (490), −0.46 (482)
FR-lex-4gr	0.002	(−0.052, 0.055)	29-21-0	−0.57 (482), 0.45 (452), 0.51 (481)
FA-alg-4gr	−0.005	(−0.026, 0.017)	20-29-1	−0.42 (559), 0.14 (577), 0.14 (570)
FA-alg-all	−0.004	(−0.022, 0.013)	10-5-35	−0.38 (559), −0.10 (591), 0.14 (570)

For example, in German topic 471 (Uhrenherstellung (Watchmaking)), the lexical stemming run substantially outscored the 4-gram run in the GenS@10 measure. The stemmer produced stems of 'uhr' and 'herstellung', with the 'uhr' (clock) stem getting higher weight from inverse document frequency, and a relevant record was retrieved at rank 4 from the 'uhr' stem matching subject terms of 'Uhr' and 'Uhrmacher'. The 4-gram approach did not match either 'Uhr' or 'Uhrmacher' (e.g. the 4-gram 'Uhre' from the query word is not a 4-gram of 'Uhr' or 'Uhrmacher') and it put a lot of weight on the less specific 'herstellung' part of the query word (several 4-gram terms) and it did not retrieve a relevant record until rank 525.

One Persian topic scored much higher in both GenS@10 and MAP using 4-grams instead of the stemmer, namely topic 559 (best Fajr film). The reason though appears to be not from 4-gramming finding better matches than the stemmer, but that the 4-gram mode did not use the stopword list. We see that topic 559 scored higher without stopping words (as per the "alg-all" line included in Table 5). The longest of the 3 Persian words in the topic title (we suspect the Persian word for 'best') was in the stopword list, perhaps inadvertently; we should investigate further.

4 Submitted Runs

For each language, we submitted 4 experimental runs in June 2008 for official assessment. In the identifiers (e.g. "otFA08tdnz"), 't', 'd' and 'n' indicate that the Title, Description and Narrative field of the topic were used (respectively), and 'e' indicates that query expansion from blind feedback on the first 3 rows was used (weight of one-half on the original query, and one-sixth each on the 3 expanded rows). The 'z' code indicates that special sampling was done, as described below. From the Description and Narrative fields for most languages, instruction words such as

Table 6: Mean Scores of Submitted Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
otDE08t	0.916	0.864	45/50	0.699	29/50	0.444	0.168	0.294
otDE08td	0.950	0.908	48/50	0.757	33/50	0.476	0.218	0.325
otDE08tde	0.911	0.849	44/50	0.688	28/50	0.496	0.210	0.357
otDE08tdz	0.957	0.914	48/50	0.758	33/50	0.476	0.171	0.247
otEN08t	0.936	0.879	45/50	0.757	33/50	0.492	0.156	0.283
otEN08td	0.956	0.905	46/50	0.764	32/50	0.460	0.199	0.290
otEN08tde	0.926	0.885	46/50	0.762	33/50	0.478	0.202	0.320
otEN08tdz	0.934	0.897	46/50	0.763	32/50	0.460	0.147	0.208
otFA08t	0.974	0.934	48/50	0.841	39/50	0.626	0.350	0.421
otFA08td	0.973	0.925	48/50	0.736	29/50	0.572	0.324	0.381
otFA08tde	0.967	0.924	48/50	0.775	33/50	0.564	0.307	0.373
(otFA08tdn)	0.970	0.916	49/50	0.712	27/50	0.560	0.273	0.335
otFA08tdnz	0.968	0.912	49/50	0.711	27/50	0.560	0.201	0.242
otFR08t	0.864	0.751	39/50	0.558	22/50	0.312	0.095	0.237
otFR08td	0.907	0.819	44/50	0.620	25/50	0.308	0.146	0.252
otFR08tde	0.823	0.741	39/50	0.552	21/50	0.284	0.121	0.250
otFR08tdz	0.883	0.803	44/50	0.617	25/50	0.308	0.110	0.192

"find", "relevant" and "document" were automatically removed (based on looking at some older topic lists, not this year's topics; this step was skipped for Persian, which was a new language this year).

Details of the submitted approaches:

- "t": Just the Title field of the topic was used. Same as the "lexstem" run of Section 2.4 for German, French and English, and same as the "algstem" run of Section 2.4 for Persian.
- "td": Same as "t" except that the Description field of the topic was additionally used.
- "tde": Same as "td" except that blind feedback (based on the first 3 rows of the "td" query) was used to expand the query.
- "tdn": Same as "td" except that the Narrative field of the topic was additionally used. (This run was not submitted.)
- "tdz": Depth-10000 sampling run based on the "td" run as described below (German, French and English only).
- "tdnz": Depth-10000 sampling run based on the "tdn" run as described below (Persian only).

Table 6 lists the mean scores for the submitted runs.

4.1 Impact of Including the Description Field

Table 7 shows the impact of including the Description field on the GenS@10 and MAP measures (and for Persian, it also shows the impact of including the Narrative field). We see large impacts on individual topics in both directions. The only statistically significant mean differences were for the Persian MAP score, for which both the Description terms and Narrative terms were detrimental.

Table 7: Impact of the Description Field on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
FR-td-t	0.069	(−0.031, 0.168)	15-13-22	0.96 (451), 0.83 (469), −0.79 (478)
DE-td-t	0.044	(−0.030, 0.117)	15-8-27	−0.98 (455), 0.73 (473), 0.79 (453)
EN-td-t	0.026	(−0.058, 0.110)	9-11-30	0.93 (473), 0.89 (487), −0.85 (459)
FA-td-t	−0.009	(−0.041, 0.022)	8-16-26	0.50 (599), 0.27 (574), −0.21 (578)
FA-tdn-td	−0.009	(−0.046, 0.029)	13-15-22	0.53 (559), −0.39 (556), −0.39 (560)
	ΔMAP			
FR-td-t	0.015	(−0.028, 0.059)	26-24-0	0.54 (451), 0.36 (469), −0.53 (452)
DE-td-t	0.031	(−0.006, 0.068)	31-19-0	0.33 (461), 0.30 (470), −0.33 (455)
EN-td-t	0.007	(−0.038, 0.052)	28-21-1	0.62 (471), 0.32 (460), −0.43 (459)
FA-td-t	−0.040	(−0.067,−0.013)	16-34-0	−0.30 (573), −0.23 (576), 0.21 (561)
FA-tdn-td	−0.045	(−0.073,−0.017)	12-38-0	0.26 (573), −0.23 (560), −0.25 (572)

Table 8: Impact of Blind Feedback on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
DE-tde-td	−0.058	(−0.104,−0.013)	3-14-33	−0.63 (459), −0.62 (478), 0.14 (458)
EN-tde-td	−0.020	(−0.061, 0.020)	9-9-32	−0.79 (475), −0.37 (456), 0.33 (463)
FR-tde-td	−0.078	(−0.134,−0.022)	5-17-28	−0.77 (463), −0.73 (452), 0.27 (466)
FA-tde-td	−0.001	(−0.033, 0.031)	11-10-29	−0.47 (574), −0.19 (566), 0.46 (559)
	ΔMAP			
DE-tde-td	0.032	(0.005, 0.059)	32-18-0	0.42 (482), 0.19 (470), −0.20 (453)
EN-tde-td	0.031	(0.004, 0.058)	33-17-0	0.42 (482), 0.26 (460), −0.14 (483)
FR-tde-td	−0.002	(−0.021, 0.017)	20-30-0	0.19 (457), −0.16 (469), −0.18 (464)
FA-tde-td	−0.007	(−0.028, 0.013)	25-25-0	0.20 (565), 0.15 (592), −0.15 (579)

4.2 Impact of Blind Feedback

Table 8 shows the impact of blind feedback on the GenS@10 and MAP measures. The results are generally consistent with our past findings that blind feedback is detrimental to GenS@10 even when it boosts MAP [11]. In particular, the mean impact was statistically significant for German in both measures (in opposite directions).

4.3 Depth-10000 Sampling

The submitted tdz or tdnz run for each language (hereinafter called the 'z run') was actually a depth probe run from sampling the td or tdn run for the language (respectively).

The base td or tdn run was retrieved to depth 10000 for each topic. The first 100 rows of the submitted z run contained the following rows of the base run in the following order:

```
1, 2, ..., 10,
20, 30, ..., 100,
200, 300, ..., 1000,
2000, 3000, ..., 10000,
15, 25, ..., 95,
150, 250, ..., 950,
1500, 2500, ..., 9500,
125, 175, ..., 975,
1250, 1750, ..., 9750.
```

Table 9: Marginal Precision of German Base-TD Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	146R, 104N, 0U	0.584	1	2.9
6-10	6, 7, ..., 10	92R, 158N, 0U	0.368	1	1.8
11-50	15, 20, ..., 50	85R, 315N, 0U	0.212	5	8.5
51-100	55, 60, ..., 100	36R, 464N, 0U	0.072	5	3.6
101-200	150, 200	4R, 96N, 0U	0.040	50	4.0
201-500	250, 300, ..., 500	6R, 294N, 0U	0.020	50	6.0
501-900	550, 600, ..., 900	2R, 398N, 0U	0.005	50	2.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	1R, 199N, 0U	0.005	500	10.0
3001-6000	3500, 4000, ..., 6000	0R, 300N, 0U	0.000	500	0.0
6001-10000	7000, 8000, ..., 10000	1R, 196N, 3X	0.005	1000	20.0

The remainder of the z run was the leftover rows from the base run until 1000 had been retrieved (rows 11, 12, 13, 14, 16, ..., 962).

This ordering (e.g. depth 10000 before depth 15) was chosen because of uncertainty of how deep the judging would be. As long as the top-37 were judged, we would have sampling to depth 10000. The extra sample points would just improve the accuracy. The z run was given highest precedence for judging. It turned out that the top-60 were judged for each topic for all 4 languages.

Tables 9, 10, 11 and 12 show the results of the sampling for each language. The columns are as follows:

- "Depth Range": The range of depths being sampled. The 11 depth ranges cover from 1 to 10000.
- "Samples": The depths of the sample points from the depth range. The samples are always uniformly spaced. They always end at the last point of the depth range. The total number of sample points (over the 11 rows of the table) adds to 60 for all 4 languages.
- "# Rel": The number of each type of item retrieved from the sample points over the 50 topics. The item type codes are R (relevant), N (non-relevant) and U (unjudged, of which there are always 0). An X is used when a sample point was not submitted because fewer than 10000 rows were retrieved for the topic (this just happened for one German topic). The sum of the item type counts is always 50 times the number of sample points for the depth range (because there are 50 topics for each language).
- "Precision": Estimated precision of the depth range (R/(R+N+U+X)).
- "Wgt": The weight of each sample point. The weight is equal to the difference in ranks between sample points, i.e. each sample point can be thought of as representing this number of rows, which is itself plus the preceding unsampled rows.
- "EstRel/Topic": Estimated number of relevant items retrieved per topic for this depth range. This is the Precision multiplied by the size of the depth range. Or equivalently, it is (R * Wgt) / 50.

Because each sample point is at the deep end of the range of rows it represents, the sampling should tend to underestimate precision for each depth range (assuming that precision tends to fall with depth, which appears to be the case for all 4 languages).

Table 13 shows the sums of the estimated number of relevant items per topic over all depth ranges in its first row. The official number of relevant items per topic for each language is listed in the second row. The final row of the table just divides the official number of relevant items

Table 10: Marginal Precision of French Base-TD Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	91R, 159N, 0U	0.364	1	1.8
6-10	6, 7, ..., 10	63R, 187N, 0U	0.252	1	1.3
11-50	15, 20, ..., 50	51R, 349N, 0U	0.128	5	5.1
51-100	55, 60, ..., 100	50R, 450N, 0U	0.100	5	5.0
101-200	150, 200	2R, 98N, 0U	0.020	50	2.0
201-500	250, 300, ..., 500	9R, 291N, 0U	0.030	50	9.0
501-900	550, 600, ..., 900	6R, 394N, 0U	0.015	50	6.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	1R, 199N, 0U	0.005	500	10.0
3001-6000	3500, 4000, ..., 6000	1R, 299N, 0U	0.003	500	10.0
6001-10000	7000, 8000, ..., 10000	0R, 200N, 0U	0.000	1000	0.0

Table 11: Marginal Precision of English Base-TD Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	137R, 113N, 0U	0.548	1	2.7
6-10	6, 7, ..., 10	93R, 157N, 0U	0.372	1	1.9
11-50	15, 20, ..., 50	96R, 304N, 0U	0.240	5	9.6
51-100	55, 60, ..., 100	75R, 425N, 0U	0.150	5	7.5
101-200	150, 200	8R, 92N, 0U	0.080	50	8.0
201-500	250, 300, ..., 500	17R, 283N, 0U	0.057	50	17.0
501-900	550, 600, ..., 900	7R, 393N, 0U	0.018	50	7.0
901-1000	950, 1000	2R, 98N, 0U	0.020	50	2.0
1001-3000	1500, 2000, ..., 3000	2R, 198N, 0U	0.010	500	20.0
3001-6000	3500, 4000, ..., 6000	2R, 298N, 0U	0.007	500	20.0
6001-10000	7000, 8000, ..., 10000	0R, 200N, 0U	0.000	1000	0.0

Table 12: Marginal Precision of Persian Base-TDN Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	145R, 105N, 0U	0.580	1	2.9
6-10	6, 7, ..., 10	135R, 115N, 0U	0.540	1	2.7
11-50	15, 20, ..., 50	136R, 264N, 0U	0.340	5	13.6
51-100	55, 60, ..., 100	145R, 355N, 0U	0.290	5	14.5
101-200	150, 200	22R, 78N, 0U	0.220	50	22.0
201-500	250, 300, ..., 500	61R, 239N, 0U	0.203	50	61.0
501-900	550, 600, ..., 900	49R, 351N, 0U	0.123	50	49.0
901-1000	950, 1000	7R, 93N, 0U	0.070	50	7.0
1001-3000	1500, 2000, ..., 3000	11R, 189N, 0U	0.055	500	110.0
3001-6000	3500, 4000, ..., 6000	9R, 291N, 0U	0.030	500	90.0
6001-10000	7000, 8000, ..., 10000	2R, 198N, 0U	0.010	1000	40.0

Table 13: Estimated Percentage of Relevant Items that are Judged, Per Topic

	DE	FR	EN	FA
Estimated Rel@10000	59.9	51.2	95.7	412.7
Official Rel/Topic	32.7	26.8	50.7	103.2
Percentage Judged	55%	52%	53%	25%

by the estimated number in the first 10000 retrieved (e.g. for German, 32.7/59.9=55%). This number should tend to be an overestimate of the percentage of all relevant items that are judged (on average per topic) because there may be relevant items that were not matched by the query in the first 10000 rows.

However, the sampling was very coarse at the deeper ranks, e.g. for German, 1 relevant item out of 200 samples in the 6001-10000 range led to an estimate of 20 relevant items per topic in this range. If the sampling had turned up 0 or 2 relevant items, a minor difference, the estimate would have been 0 or 40 relevant items per topic in this range, leading to a substantially different sum (39.9 or 79.9 instead of 59.9). We should compute confidence intervals for these estimates, but have not yet done so. Also, there is a lot of variance across topics, which we have not yet analyzed.

These preliminary estimates of judging coverage for the CLEF 2008 collections (55% for German, 52% for French, 53% for English, 25% for Persian) tend to be lower than the estimates we produced for the CLEF 2007 collections last year [16] (55% for Czech, 69% for Bulgarian, 83% for Hungarian) or the estimates we produced for the NTCIR-6 collections (58% for Chinese, 78% for Japanese, 100% for Korean) [15]. The German, French and English estimates are higher than the estimates we produced for the TREC 2006 Legal and Terabyte collections using a similar approach (18% for TREC Legal and 36% for TREC Terabyte) [12], while the Persian estimate is in the same ballpark as these (much larger) TREC 2006 collections.

For Persian, the topics appear to have been relatively broad (more relevant documents per topic on average) which led to the judging coverage being relatively shallow (based on the sampling experiment). It is not clear however whether these are factors in the unusual results we found for Persian (e.g. normally the Description and Narrative terms increase retrieval scores instead of decrease, and normally we see more impact from stemming on at least some individual topics).

The incompleteness results for German, French and English are similar to what [18] found for depth-100 pooling on the old TREC collections of approximately 500,000 documents: "it is likely that at best 50%-70% of the relevant documents have been found; most of these unjudged relevant documents are for the 10 or so queries that already have the most known answers."

Fortunately, [18] also found for such test collections that "overall they do indeed lead to reliable results." (We can also confirm that we have gained a lot of insights from the CLEF test collections over the years, such as from the topic analyses in [13].)

References

[1] Samir Abdou and Jacques Savoy. Considérations sur l'évaluation de la robustesse en recherche d'information. *CORIA 2007*.

[2] Harr Chen and David R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. *SIGIR 2006*, pp. 429-436.

[3] Cross-Language Evaluation Forum web site. http://www.clef-campaign.org/

[4] Andrew Hodgson. Converting the Fulcrum Search Engine to Unicode. *Sixteenth International Unicode Conference*, 2000.

[5] NTCIR (NII-NACSIS Test Collection for IR Systems) Home Page. http://research.nii.ac.jp/~ntcadm/index-en.html

[6] M. F. Porter. Snowball: A language for stemming algorithms. October 2001. http://snowball.tartarus.org/texts/introduction.html

[7] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. *Proceedings of TREC-3*, 1995.

[8] Jacques Savoy. CLEF and Multilingual information retrieval resource page. http://www.unine.ch/info/clef/

[9] Text REtrieval Conference (TREC) Home Page. http://trec.nist.gov/

[10] Stephen Tomlinson. Comparing the Robustness of Expansion Techniques and Retrieval Measures. *Working Notes for the CLEF 2006 Workshop.*

[11] Stephen Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. *SIGIR 2006*, pp. 705-706.

[12] Stephen Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track. *Proceedings of TREC 2006.*

[13] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer™ at CLEF 2005. *Working Notes for the CLEF 2005 Workshop.*

[14] Stephen Tomlinson. Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer™ at CLEF 2003. *Working Notes for the CLEF 2003 Workshop.*

[15] Stephen Tomlinson. Sampling Precision to Depth 9000: Evaluation Experiments at NTCIR-6. *Proceedings of NTCIR-6*, 2007.

[16] Stephen Tomlinson. Sampling Precision to Depth 10000: Evaluation Experiments at CLEF 2007. *Working Notes for the CLEF 2007 Workshop.*

[17] Ellen M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. *Proceedings of TREC 2004.*

[18] Justin Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? *SIGIR'98*, pp. 307-314.