European Web Retrieval Experiments with Hummingbird SearchServer$^{\text{TM}}$ at CLEF 2005

Stephen Tomlinson
Hummingbird
Ottawa, Ontario, Canada
stephen.tomlinson@hummingbird.com
http://www.hummingbird.com/

August 22, 2005

Abstract

Hummingbird participated in the mixed monolingual retrieval task of the WebCLEF Track of the Cross-Language Evaluation Forum (CLEF) 2005. In this task, the system was given 547 known-item queries from 11 languages (134 Spanish, 121 English, 59 Dutch, 59 Portuguese, 57 German, 35 Hungarian, 30 Danish, 30 Russian, 16 Greek, 5 Icelandic and 1 French). The goal was to find the desired page in the 82GB EuroGOV collection (3.4 million pages crawled from government sites of 27 European domains). We experimented with different techniques for web retrieval and analyzed the differences between them. We defined a new measure, First Relevant Score (FRS), to facilitate per-topic analysis, and we focused on analyzing Greek, Danish and Icelandic topics. We found that stopword processing was more important than anticipated, perhaps because words common in one language may tend to be overweighted by inverse document frequency in a mixed language collection. Extra weight on the document title helped significantly, and extra weight on less deep urls significantly helped home page queries. Stemming was of neutral impact on average, but could make a substantial difference for individual queries.

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

Greek Retrieval, Danish Retrieval, Icelandic Retrieval, First Relevant Score, Per-Topic Analysis

1 Introduction

Hummingbird SearchServer[1] is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other Hummingbird products for the enterprise.

[1] SearchServer$^{\text{TM}}$, SearchSQL$^{\text{TM}}$and Intuitive Searching$^{\text{TM}}$ are trademarks of Hummingbird Ltd. All other copyrights, trademarks and tradenames are the property of their respective owners.

SearchServer works in Unicode internally [3] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (CLEF [2], NTCIR [4] and TREC [10]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This (draft) paper describes experimental work with SearchServer for the task of finding known home or named pages in 11 European languages (Spanish, English, Dutch, Portuguese, German, Hungarian, Danish, Russian, Greek, Icelandic and French) using the WebCLEF 2005 test collection.

2 Methodology

For the submitted runs in June 2005, SearchServer experimental development build 7.0.0.707 was used.

2.1 Data

The collection to be searched was the EuroGOV collection. It consisted of 3,589,502 pages crawled from government sites of 27 European domains. Uncompressed, it was 88,062,007,676 bytes (82.0 GB). The average document size was 24,533 bytes. More details on this collection are in [8]. Note that we only indexed 3,417,463 of the pages because the organizers provided a "blacklist" of 172,039 pages to omit (primarily binary documents).

For the mixed monolingual task, there were 547 known-item queries from 11 different languages (134 Spanish, 121 English, 59 Dutch, 59 Portuguese, 57 German, 35 Hungarian, 30 Danish, 30 Russian, 16 Greek, 5 Icelandic and 1 French). Of these, 345 were named page queries and 242 were home page queries. More details on the mixed monolingual task are in the track overview paper [9].

2.2 Indexing

Our indexing approach was based on the approach we used for TREC Web tasks the previous three years (described in detail in [12]). Briefly, in addition to full-text indexing, the custom text reader cTREC populated particular columns such as TITLE (if any), URL, URL_TYPE and URL_DEPTH. The URL_TYPE was set to ROOT, SUBROOT, PATH or FILE, based on the convention which worked well in TREC 2001 for the Twente/TNO group [15] on the entry page finding task (also known as the home page finding task). The URL_DEPTH was set to a term indicating the depth of the page in the site. Table 1 contains URL types and depths for example URLs. The exact rules we used are given in [12].

WebCLEF required a few indexing enhancements compared to TREC. In particular, it wouldn't suffice to assume all the pages were in the ASCII character set. We added a /cs option to our cTREC text reader which used the first recognized 'charset' specification in the page (e.g. from the meta http-equiv tag) to indicate from which character set to convert the page to Unicode (Win_1252 was assumed if no charset was specified).

For the baseline task, in which the system was not to make use of any of the topic metadata such as the specified language of the query, we still indexed with English stopwords (even though the majority of the documents were in other languages). We treated the apostrophe as a term separator (which we normally do for languages other than English, but in this collection, it was also a separator for English). No accents were indexed. English stemming was used on the table, but SearchServer also indexed all the surface forms (after Unicode normalizations such as case normalization), and the baseline runs just searched the surface forms, not the stems.

For 2 of our submitted runs, we labelled the runs as making use of the topic and page language metadata (which were always the same in the mixed monolingual task) along with the page's domain. For these runs, we created a set of language-specific indexes (one for each of the 11 query languages) which used a stemmer and stopfile for that language (for English and Icelandic,

Table 1: Examples of URL Type and Depth Values

URL	Type	Depth	Depth Term
http://nasa.gov/	ROOT	1	URLDEPTHA
http://www.nasa.gov/	ROOT	1	URLDEPTHA
http://jpl.nasa.gov/	ROOT	2	URLDEPTHAB
http://fred.jpl.nasa.gov/	ROOT	3	URLDEPTHABC
http://nasa.gov/jpl/	SUBROOT	2	URLDEPTHAB
http://nasa.gov/jpl/fred/	PATH	3	URLDEPTHABC
http://nasa.gov/index.html	ROOT	1	URLDEPTHA
http://nasa.gov/fred.html	FILE	2	URLDEPTHAB

we actually used the original baseline index, which had English stems and stopwords). For some of the languages, because we were close to the submission deadline, we also skipped indexing some of the domains to save time (e.g. for Greek, just the ‘gr’ and ‘eu.int’ subsets of EuroGOV were included because it was known all the results were in the ‘gr’ domain) which would have a (probably minor) effect on the inverse document frequencies (minor especially since we always included the ‘eu.int’ subset in each index). For 9 of the languages (Danish, Dutch, English, French, German, Greek, Portuguese, Russian and Spanish), the lexical stemmer in SearchServer (based on internal stemming component 3.7.0.15) was used. For Hungarian, the Neuchatel stemmer [7] was used (see our companion ad hoc retrieval paper [11] for details). For Icelandic, we used the English index as previously mentioned. For Greek and Russian, we additionally enabled indexing of a few accents because the stemmer was accent-sensitive. When processing queries for these runs, the query was directed to the index for the specified language.

2.3 Searching

We executed 7 runs in June 2005, though only 5 were allowed to be submitted. All 7 are described here. The first 4 runs were ‘baseline’ runs which did not use the topic metadata. The other 3 runs made use of the topic metadata (in particular, the domain, and for the last 2 runs, also the language).

humWC05none: This run was a plain content search of the baseline table. No inflections were used. This run was the analog of the “none” runs described in our ad hoc retrieval paper [11]. It used the ‘2:3’ relevance method and document length normalization (SET RELEVANCE_DLEN_IMP 500). The IS_ABOUT predicate was used instead of the CONTAINS predicate (and hence the VECTOR_GENERATOR was set to blank to disable inflections instead of the TERM_GENERATOR), but the relevance calculation was the same. (This run was not submitted.)

humWC05p run: This submitted run was the same as humWC05none except that it put additional weight on matches in the title, url, first heading and some meta tags, including extra weight on matching the query as a phrase in these fields. Below is an example SearchSQL query. The searches on the ALL_PROPS column (which contained a copy of the title, url, etc. as described in [12]) are the difference from the humWC05none run. Note that the FT_TEXT column indexed the content and also all of the non-content fields except for the URL. More details of the syntax are explained in [13]. This run used the same approach as the TREC 2004 humW04pl run except that linguistic inflections were disabled.

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM EGOV
WHERE
 (ALL_PROPS CONTAINS 'Giuseppe Medici' WEIGHT 1) OR
 (ALL_PROPS IS_ABOUT 'Giuseppe Medici' WEIGHT 1) OR
 (FT_TEXT IS_ABOUT 'Giuseppe Medici' WEIGHT 10)
ORDER BY REL DESC;
```

humWC05dp run: This submitted run was the same as humWC05p except that it put additional weight on urls of depth 4 or less (but not on the url type, though url types were still listed with weight 0 as a way to prevent urls of depth greater than 4 from being excluded). Less deep urls also received higher weight from inverse document frequency because (presumably) they are less common. This run used the same approach as the TREC 2004 humW04dpl run except that linguistic inflections were disabled. Below is an example WHERE clause:

```
WHERE
((ALL_PROPS CONTAINS 'Giuseppe Medici' WEIGHT 1) OR
 (ALL_PROPS IS_ABOUT 'Giuseppe Medici' WEIGHT 1) OR
 (FT_TEXT IS_ABOUT 'Giuseppe Medici' WEIGHT 10)
) AND (
 (URL_TYPE CONTAINS 'ROOT' WEIGHT 0) OR
 (URL_TYPE CONTAINS 'SUBROOT' WEIGHT 0) OR
 (URL_TYPE CONTAINS 'PATH' WEIGHT 0) OR
 (URL_TYPE CONTAINS 'FILE' WEIGHT 0) OR
 (URL_DEPTH CONTAINS 'URLDEPTHA' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHAB' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHABC' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHABCD' WEIGHT 5) )
```

humWC05rdp run: This submitted run was the same as humWC05dp except that it put additional weight on the url type. This run used the same approach as the TREC 2004 humW04rdpl run except that linguistic inflections were disabled. Below is an example WHERE clause:

```
WHERE
((ALL_PROPS CONTAINS 'Giuseppe Medici' WEIGHT 1) OR
 (ALL_PROPS IS_ABOUT 'Giuseppe Medici' WEIGHT 1) OR
 (FT_TEXT IS_ABOUT 'Giuseppe Medici' WEIGHT 10)
) AND (
 (URL_TYPE CONTAINS 'ROOT' WEIGHT 10) OR
 (URL_TYPE CONTAINS 'SUBROOT' WEIGHT 10) OR
 (URL_TYPE CONTAINS 'PATH' WEIGHT 10) OR
 (URL_TYPE CONTAINS 'FILE' WEIGHT 0) OR
 (URL_DEPTH CONTAINS 'URLDEPTHA' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHAB' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHABC' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHABCD' WEIGHT 5) )
```

humWC05dpD0 run: This run was the same as humWC05dp except that the domain information of the topic metadata was used to restrict the search to the specified domain. Below is an example of the domain filter added to the WHERE clause for a case in which the page was known to be in the 'it' domain (which implied the DOCNO would contain 'Eit'). This run was not submitted.

```
AND (DOCNO CONTAINS 'Eit' WEIGHT 0)
```

humWC05dpD run: This submitted run was the same as humWC05dpD0 except that the language information of the topic metadata was used to direct the search to the table for the specified language (i.e. the WHERE clause was the same as for humWC05dpD0, but the FROM clause specified a different table). Inflections were still not used.

humWC05dplD run: This submitted run was the same as humWC05dpD except that the content and title searches included linguistic expansion from language-specific stemming (this was done with SET VECTOR_GENERATOR 'word!ftelp/inflect'; note that /decompound (applicable to Dutch and German) is implied for /inflect with SET VECTOR_GENERATOR, unlike with SET TERM_GENERATOR).

2.4 Evaluation Measures

If one wishes to focus on just the first relevant document, the traditional measure is "Reciprocal Rank" (RR). For a topic, it 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.

An experimental measure introduced in this paper (along with the companion ad hoc retrieval paper [11]) is "First Relevant Score" (denoted "FRS"). Like reciprocal rank, it is based on just the rank of the first relevant retrieved for a topic, but it is better suited to per-topic analysis. FRS 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. Like reciprocal rank, finding the first relevant at rank 1 produces a score of 1.0. At rank 2, FRS is just 7 points lower (0.93), whereas RR is 50 points lower (0.50). At rank 3, FRS is another 7 points lower (0.86), whereas RR is 17 points lower (0.33). At rank 10, FRS is 0.50, whereas RR is 0.10. FRS is greater than RR for ranks 2 to 52 and lower for ranks 53 and beyond. A possible interpretation of FRS is that it may be an indicator of the percentage of potential result list reading the system saved the user to get to the first relevant, assuming that users are less and less likely to continue reading as they get deeper into the result list.

"Success@n" is the percentage of topics for which at least one relevant document was returned in the first n rows. Like the other first relevant measures, this measure hides a lot of retrieval differences (particularly in recall), but it is more intuitive and may be an indicator of a user's impression of a method's robustness across topics. This paper lists Success@1, Success@5 and Success@10.

2.5 Per-Topic Tables

The 7 runs allow us to isolate 6 'web techniques' which are denoted as follows:

- 'p' (extra weight for phrases in the Title and other properties plus extra weight for vector search on properties): The humWC05p score minus the humWC05none score.
- 'd' (modest extra weight for less deep urls): The humWC05dp score minus the humWC05p score.
- 'r' (strong extra weight for urls of root, subroot or path types): The humWC05rdp score minus the humWC05dp score.
- 'o' (domain filtering): The humWC05dpD0 score minus the humWC05dp score.
- 's' (stopwords specific to the language and possibly accent-indexing and inverse document frequency changes): The humWC05dpD score minus the humWC05dpD0 score.
- 'l' (linguistic expansion from stemming): The humWC05dplD score minus the humWC05dpD score.

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

- "Expt" specifies the experiment. It starts with one of the above 6 web techniques, followed by 'NP' for named page queries or 'HP' for home page queries, optionally followed by the language code.
- "ΔFRS" is the difference of the (mean) first relevant score of the two runs being compared.

Table 2: Mean Scores of Submitted WebCLEF Runs

Run	FRS	Success@1	Success@5	Success@10	MRR
humWC05dplD	0.635	212/547 (39%)	315/547 (58%)	356/547 (65%)	0.478
humWC05dpD	0.627	204/547 (37%)	314/547 (57%)	353/547 (65%)	0.471
(humWC05dpD0)	0.603	197/547 (36%)	303/547 (55%)	343/547 (63%)	0.449
humWC05rdp	0.589	195/547 (36%)	293/547 (54%)	330/547 (60%)	0.441
humWC05dp	0.583	190/547 (35%)	292/547 (53%)	327/547 (60%)	0.433
humWC05p	0.559	182/547 (33%)	276/547 (50%)	318/547 (58%)	0.415
(humWC05none)	0.513	152/547 (28%)	253/547 (46%)	284/547 (52%)	0.365
NP: dplD	0.726	139/305 (46%)	204/305 (67%)	229/305 (75%)	0.560
NP: dpD	0.720	142/305 (47%)	207/305 (68%)	228/305 (75%)	0.571
NP: dpD0	0.698	141/305 (46%)	202/305 (66%)	223/305 (73%)	0.552
NP: rdp	0.662	132/305 (43%)	187/305 (61%)	206/305 (68%)	0.517
NP: dp	0.669	134/305 (44%)	192/305 (63%)	210/305 (69%)	0.526
NP: p	0.669	133/305 (44%)	193/305 (63%)	212/305 (70%)	0.527
NP: none	0.648	119/305 (39%)	187/305 (61%)	203/305 (67%)	0.492
HP: dplD	0.521	73/242 (30%)	111/242 (46%)	127/242 (52%)	0.375
HP: dpD	0.509	62/242 (26%)	107/242 (44%)	125/242 (52%)	0.345
HP: dpD0	0.484	56/242 (23%)	101/242 (42%)	120/242 (50%)	0.319
HP: rdp	0.497	63/242 (26%)	106/242 (44%)	124/242 (51%)	0.345
HP: dp	0.474	56/242 (23%)	100/242 (41%)	117/242 (48%)	0.317
HP: p	0.420	49/242 (20%)	83/242 (34%)	106/242 (44%)	0.275
HP: none	0.343	33/242 (14%)	66/242 (27%)	81/242 (33%)	0.205

- "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 in the particular experiment.

- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets (the topic numbers range from 1 to 547). 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).

3 Results of Web Experiments

Table 2 lists the mean scores of the 5 submitted runs (and the 2 other diagnostic runs in brackets). It also lists the mean scores over just named page (NP) and home page (HP) queries.

Table 3 isolates the differences in 'first relevant score' (FRS) between the runs of Table 2.

- The 'p' technique (extra weight for phrases in the Title and other properties plus extra weight for vector search on properties) was of statistically significant benefit for both named pages and home pages, which is consistent with our TREC results [14] except that the benefit was larger at TREC.

Table 3: Impact of Web Techniques on First Relevant Score

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
o-NP	0.029	(0.015, 0.042)	44-0-261	1.00 (285), 0.87 (292), 0.00 (289)
s-NP	0.022	(0.007, 0.037)	43-24-238	−0.88 (527), 0.76 (477), 0.87 (116)
p-NP	0.021	(0.007, 0.035)	65-28-212	−0.83 (292), −0.64 (477), 0.59 (415)
l-NP	0.006	(−0.014, 0.025)	47-59-199	1.00 (112), 0.95 (402), −0.78 (157)
d-NP	0.000	(−0.005, 0.005)	24-37-244	0.40 (377), 0.17 (524), −0.14 (423)
r-NP	−0.008	(−0.018, 0.003)	18-49-238	−0.87 (469), −0.46 (528), 0.68 (418)
p-HP	0.077	(0.050, 0.105)	82-19-141	1.00 (101), 0.98 (313), −0.92 (435)
d-HP	0.054	(0.032, 0.075)	64-20-158	1.00 (453), 0.91 (52), −0.40 (290)
s-HP	0.025	(0.005, 0.044)	53-21-168	0.91 (39), −0.76 (346), −0.79 (20)
r-HP	0.023	(−0.009, 0.054)	53-48-141	−1.00 (148), −0.93 (246), 0.92 (32)
l-HP	0.012	(−0.011, 0.036)	41-50-151	0.96 (123), 0.93 (124), −0.68 (324)
o-HP	0.010	(0.003, 0.017)	22-0-220	0.43 (432), 0.40 (507), 0.00 (546)

- The ‘d’ technique (modest extra weight for less deep urls) was of statistically significant benefit for home pages and neutral on average for named pages, which is consistent with our TREC results except that the benefit for home pages was larger at TREC.
- The ‘r’ technique (strong extra weight for urls of root, subroot or path types) was less detrimental than we expected for named pages and less helpful than we expected for home pages compared to our TREC results.
- The ‘o’ technique (domain filtering), as expected, never caused the score to go down on any topic (as the ‘vs.’ column shows) because it just included rows from the known domain. But the benefit was not large on average, so apparently the unfiltered queries usually were not confused much by the extra domains.
- The ‘s’ technique (stopwords specific to the language and possibly accent-indexing and inverse document frequency changes) was a surprise in that it led to a statistically significant benefit for both named pages and home pages. We look at this more below.
- The ‘l’ technique (linguistic expansion from stemming) was of neutral impact on average, but it could make a substantial difference for individual queries as we will see below.

In the sections that follow, we focus on Greek, Danish and Icelandic because this is the first time we have had judged test collections for these languages. In partciular, we focus on the impact of the ‘s’ and ‘l’ techniques, i.e. the impacts of stopwords (and accents) and stemming. For English, we compare the scores on our own contributed topics to the other English topics. The last section lists the per-topic tables for the remaining languages in descending order by number of topics, for future reference.

Table 4: Mean Scores of WebCLEF Runs on Greek Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dplD-NP-EL	0.536	3/11 (27%)	5/11 (45%)	6/11 (55%)	0.363
dpD0-NP-EL	0.442	3/11 (27%)	5/11 (45%)	5/11 (45%)	0.316
dpD-NP-EL	0.398	1/11 (9%)	4/11 (36%)	5/11 (45%)	0.206
dp-NP-EL	0.306	3/11 (27%)	3/11 (27%)	3/11 (27%)	0.279
rdp-NP-EL	0.297	2/11 (18%)	3/11 (27%)	3/11 (27%)	0.233
p-NP-EL	0.291	3/11 (27%)	3/11 (27%)	3/11 (27%)	0.277
none-NP-EL	0.287	2/11 (18%)	3/11 (27%)	3/11 (27%)	0.232
dpD0-HP-EL	0.657	2/5 (40%)	3/5 (60%)	3/5 (60%)	0.483
dplD-HP-EL	0.571	2/5 (40%)	3/5 (60%)	3/5 (60%)	0.467
rdp-HP-EL	0.571	2/5 (40%)	3/5 (60%)	3/5 (60%)	0.467
dp-HP-EL	0.571	2/5 (40%)	3/5 (60%)	3/5 (60%)	0.467
dpD-HP-EL	0.532	1/5 (20%)	3/5 (60%)	3/5 (60%)	0.340
p-HP-EL	0.480	1/5 (20%)	2/5 (40%)	3/5 (60%)	0.289
none-HP-EL	0.430	1/5 (20%)	2/5 (40%)	2/5 (40%)	0.278

Table 5: Impact of Web Techniques on First Relevant Score, Greek Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
l-NP-EL	0.138	(−0.056, 0.333)	5-2-4	1.00 (112), 0.34 (395), −0.15 (403)
o-NP-EL	0.136	(−0.015, 0.288)	3-0-8	0.74 (403), 0.40 (395), 0.00 (266)
d-NP-EL	0.015	(−0.016, 0.046)	1-0-10	0.17 (524), 0.00 (151), 0.00 (266)
p-NP-EL	0.004	(−0.012, 0.020)	1-1-9	0.07 (184), 0.00 (151), −0.03 (524)
r-NP-EL	−0.009	(−0.024, 0.005)	0-2-9	−0.07 (184), −0.03 (524), 0.00 (266)
s-NP-EL	−0.045	(−0.117, 0.028)	1-4-6	−0.34 (395), −0.14 (498), 0.13 (445)
d-HP-EL	0.092	(−0.092, 0.276)	1-0-4	0.46 (366), 0.00 (271), 0.00 (25)
o-HP-EL	0.086	(−0.086, 0.258)	1-0-4	0.43 (432), 0.00 (366), 0.00 (25)
p-HP-EL	0.050	(−0.050, 0.150)	1-0-4	0.25 (366), 0.00 (271), 0.00 (25)
l-HP-EL	0.039	(−0.012, 0.090)	2-0-3	0.12 (366), 0.07 (271), 0.00 (25)
r-HP-EL	0.000	n/a	0-0-5	0.00 (271), 0.00 (366), 0.00 (25)
s-HP-EL	−0.125	(−0.316, 0.066)	1-2-2	−0.43 (432), −0.26 (366), 0.07 (271)

3.1 Greek Retrieval

Table 4 lists the mean scores for the 11 Greek named page queries and 5 Greek home page queries. The top-scoring runs used stemming (run humWC05dplD) or disabled accent-indexing (run humWC05dplD0). The run with accent-indexing and not stemming (humWC05dpD) did not score as highly on average. Table 5 shows that the ‘l’ technique (stemming, i.e. the dplD score minus the dpD score) was positive on average, while the ‘s’ factor (the dpD score minus the dpD0 score, primarily isolating the impact of stopwords specific to the language, including specifying accent-indexing in the Greek case) was negative, and it lists the topics most affected by each technique in each direction, which we examine below. (In the below topic-analysis, the translations are based partly on the official topic translations and partly on the online Greek-to-English translation service at [1]).

WC0112: Table 5 shows that the biggest impact of Greek stemming was on topic 112 (Πλήρης λίστα των υπουργών και υφυπουργών όλων υπουργείων της Ελληνικής κυβέρνησης (List of ministers and deputy ministers for all the ministries of the Greek government)). The desired page was not retrieved in the top-50 without inflecting because the key query terms were plurals (υπουργών (ministers), υφυπουργών (undersecretaries), υπουργείων (ministries)) while the desired page just contained singular forms (Υπουργός (Minister), Υφυπουργός (Undersecretary), Υπουργείο (Ministry)).

WC0395: Table 5 shows that the next biggest impact of Greek stemming was on topic 395 (O ´Ελληνας πρωθυπουργός και το μήνυμά του (The Greek Prime Minister and his message)). With stemming, the first relevant was found at rank 13 instead of 39, a 34 point increase in FRS (in the reciprocal rank measure, this would just be a 5 point increase). Without stemming, the only matching word was του (his), which probably should have been a stopword. With stemming, the query word πρωθυπουργός (Prime Minister) matched the document's variant (Πρωθυπουργού). Because we enabled indexing of Greek accents for our lexical Greek stemmer, the query word μήνυμά (message) did not match the document form Μήνυμα (which did not include an accent on the last character; the first letter is just an lowercase-uppercase difference which all runs handled by normalizing Unicode to uppercase). Note that the humWC05dpD0 run did match Μήνυμα because accent-indexing was not enabled for this run; presumably this is why the s-NP-EL line of Table 5 shows that switching to the Greek-specific stopfile (which enabled accent indexing) decreased FRS 34 points for this topic. For most languages, our lexical stemmers are accent-insensitive; we should investigate doing the same for Greek.

WC0432: Table 5 shows that the biggest impact of switching to the Greek-specific stopfile was a detrimental impact on topic 432 (Είσοδος Ελληνικής ιστοσελίδας για τη συνέλευση για το μέλλον της Ευρώπης (Greek home page of the convention for the future of Europe)). The desired page was found at rank 12 without accent-indexing but was not retrieved in the top-50 with accent-indexing. The humWC05dpD0 run matched the document title terms which were in uppercase and did not have accents, particularly ΣΥΝΕΛΕΥΣΗ (ASSEMBLY), ΜΕΛΛΟΝ (FUTURE) and ΕΥΡΩΠΗΣ (EUROPE). (The corresponding query words had accents: συνέλευση (assembly), μέλλον (future) and Ευρώπης (Europe)). This issue would presumably impair the 'p' web technique (extra weight on properties such as the title) because title words are often in uppercase and apparently in Greek uppercase words often omit the accents. (Incidentally, the o-HP-EL line of Table 5 shows that domain filtering (restricting to the .gr domain) was useful for this query; without it, even without accent-indexing, the retrieved pages were mostly from the .eu.int domain.)

WC0445: Table 5 shows that the biggest positive impact of switching to the Greek-specific stopfile was on topic 445 (Πληροφορίες επικοινωνίας όλων των υπουργείων της Ελληνικής κυβέρνησης (Contact information of all the ministries of the Greek government)). The reason seems to be that the non-content words in the query (such as των (of) and της (her)) generated spurrious matches in the humWC05dpD0 run (which did not use Greek-specific stopwords), pushing down the desired page from rank 28 to beyond the top-50. Normally, common words have little effect on the ranking because they have a low inverse document frequency (idf), but in this mixed language collection, common words in the Greek documents are still fairly uncommon overall, and hence get relatively more weight. This topic illustrates why stopword processing may be of more importance in mixed language collections than in single language collections.

Even though there were just 16 Greek topics, with careful experimental setup and detailed per-topic analysis, we learned a lot about Greek web search in a mixed language collection. Stemming can be quite helpful, accent mismatches are common (especially in the important Title field of web documents), and stopwords common in one language may be over-weighted in a mixed language collection by traditional idf formulations.

Table 6: Mean Scores of WebCLEF Runs on Danish Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dplD-NP-DA	0.807	12/19 (63%)	14/19 (74%)	15/19 (79%)	0.693
dpD-NP-DA	0.798	11/19 (58%)	15/19 (79%)	16/19 (84%)	0.661
dpD0-NP-DA	0.759	11/19 (58%)	14/19 (74%)	15/19 (79%)	0.632
p-NP-DA	0.758	10/19 (53%)	13/19 (68%)	15/19 (79%)	0.616
dp-NP-DA	0.754	11/19 (58%)	13/19 (68%)	15/19 (79%)	0.629
rdp-NP-DA	0.743	10/19 (53%)	13/19 (68%)	15/19 (79%)	0.593
none-NP-DA	0.704	9/19 (47%)	12/19 (63%)	14/19 (74%)	0.550
rdp-HP-DA	0.336	1/11 (9%)	2/11 (18%)	4/11 (36%)	0.158
dpD-HP-DA	0.320	1/11 (9%)	2/11 (18%)	4/11 (36%)	0.147
dpD0-HP-DA	0.310	0/11 (0%)	2/11 (18%)	3/11 (27%)	0.108
dp-HP-DA	0.310	0/11 (0%)	2/11 (18%)	3/11 (27%)	0.108
dplD-HP-DA	0.301	1/11 (9%)	1/11 (9%)	3/11 (27%)	0.135
p-HP-DA	0.242	0/11 (0%)	1/11 (9%)	2/11 (18%)	0.067
none-HP-DA	0.163	0/11 (0%)	1/11 (9%)	1/11 (9%)	0.052

Table 7: Impact of Web Techniques on First Relevant Score, Danish Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
p-NP-DA	0.053	(0.003, 0.104)	7-1-11	0.39 (311), 0.25 (329), −0.07 (264)
s-NP-DA	0.040	(−0.037, 0.116)	2-1-16	0.71 (233), 0.12 (329), −0.08 (58)
l-NP-DA	0.008	(−0.050, 0.067)	4-1-14	−0.43 (329), 0.13 (311), 0.25 (219)
o-NP-DA	0.005	(−0.002, 0.013)	2-0-17	0.05 (329), 0.04 (58), 0.00 (481)
d-NP-DA	−0.004	(−0.035, 0.027)	2-4-13	0.14 (454), 0.14 (329), −0.14 (58)
r-NP-DA	−0.011	(−0.027, 0.005)	0-3-16	−0.14 (211), −0.05 (329), 0.00 (232)
p-HP-DA	0.079	(0.013, 0.144)	6-0-5	0.27 (80), 0.23 (48), 0.00 (429)
d-HP-DA	0.068	(−0.076, 0.211)	2-1-8	0.78 (286), 0.02 (392), −0.05 (53)
r-HP-DA	0.027	(−0.013, 0.066)	3-0-8	0.21 (385), 0.07 (286), 0.00 (429)
s-HP-DA	0.011	(−0.034, 0.055)	4-3-4	0.15 (80), 0.07 (286), −0.13 (48)
o-HP-DA	0.000	n/a	0-0-11	0.00 (317), 0.00 (53), 0.00 (429)
l-HP-DA	−0.019	(−0.069, 0.030)	3-2-6	−0.21 (317), −0.13 (48), 0.08 (392)

3.2 Danish Retrieval

WC0233: Table 7 shows that the biggest impact of switching to the Danish-specific stopfile was a 71 point increase in FRS on topic 233 (presserum europæiske kantor for bekæmpelse af svig (press room of the european anti fraud office)). Without having ‘af’ as a stopword, the first relevant rank fell from 2 to 21. This appears to be a similar finding to Greek topic WC0445 in that a common word in one language was uncommon enough in the mixed language collection to be assigned a high enough inverse document frequency to cause trouble. (Our Danish stoplist was based on Porter’s [5].) Incidentally, with stemming enabled, the rank increased from 2 to 1 for this topic, in part because of an extra ‘bekaempelse’ match in the meta keywords and also from an extra ‘Europaeiske’ match in body. It’s good to see that the SearchServer stemmer handled the æ vs. ae variation of Danish (the query words used the one-character ligature (æ) while the document words used two letters (‘a’ and ‘e’)).

WC0392: Another interesting Danish stemming case was on topic 392 (Rigsombudsmanden i Grønland (the high commissioner of greenland)). With stemming, the rank of the desired page increased from 24 to 19. The extra matches from stemming were ‘Rigsombudsmand’ and ‘Groenland' (the latter occurred in the filenames of img tags, which we indexed). Again, it's good to see that the SearchServer stemmer matched the query form using the Danish o with stroke (ø) with the two-letter variant ('oe').

WC0317: On topic 317 (økologisk landbrug i europa (organic farming in europe)), the rank of the desired page actually fell from 4 to 8 with stemming, even though the additional matches of 'okologisk' (in the meta keywords) and 'landbrugets' look proper. (As an aside, the compound 'landbrugspolitik' was not matched; we're unsure in general how common compound words are in Danish.) The relevance scores of the top documents were close together for this topic, so the fall in rank appears to be a chance result. Note that the cTREC text reader used for these experiments did not normalize the html entity reference '&Oslash;' to Ø (or most other entity references for that matter, which may have impaired the overall results for some languages). It's good to see that the SearchServer stemmer matched the query form using the Danish o with stroke (ø) with the one-letter variant ('o').

Table 8: Mean Scores of WebCLEF Runs on Icelandic Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dpD-NP-IS	0.745	2/4 (50%)	2/4 (50%)	3/4 (75%)	0.550
dpD0-NP-IS	0.745	2/4 (50%)	2/4 (50%)	3/4 (75%)	0.550
dp-NP-IS	0.731	2/4 (50%)	2/4 (50%)	3/4 (75%)	0.548
dplD-NP-IS	0.727	1/4 (25%)	2/4 (50%)	3/4 (75%)	0.425
p-NP-IS	0.727	2/4 (50%)	2/4 (50%)	3/4 (75%)	0.546
rdp-NP-IS	0.670	2/4 (50%)	2/4 (50%)	2/4 (50%)	0.534
none-NP-IS	0.629	2/4 (50%)	2/4 (50%)	2/4 (50%)	0.527
rdp-HP-IS	0.500	0/1 (0%)	0/1 (0%)	1/1 (100%)	0.100
dplD-HP-IS	0.271	0/1 (0%)	0/1 (0%)	0/1 (0%)	0.056
dpD-HP-IS	0.271	0/1 (0%)	0/1 (0%)	0/1 (0%)	0.056
dpD0-HP-IS	0.271	0/1 (0%)	0/1 (0%)	0/1 (0%)	0.056
dp-HP-IS	0.271	0/1 (0%)	0/1 (0%)	0/1 (0%)	0.056
p-HP-IS	0.271	0/1 (0%)	0/1 (0%)	0/1 (0%)	0.056
none-HP-IS	0.232	0/1 (0%)	0/1 (0%)	0/1 (0%)	0.050

Table 9: Impact of Web Techniques on First Relevant Score, Icelandic Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
p-NP-IS	0.098	(−0.018, 0.215)	2-0-2	0.22 (456), 0.17 (46), 0.00 (6)
o-NP-IS	0.014	(−0.015, 0.043)	1-0-3	0.06 (46), 0.00 (456), 0.00 (6)
d-NP-IS	0.004	(−0.025, 0.034)	1-1-2	0.04 (456), 0.00 (488), −0.03 (46)
s-NP-IS	0.000	n/a	0-0-4	0.00 (46), 0.00 (456), 0.00 (6)
l-NP-IS	−0.019	(−0.056, 0.019)	0-1-3	−0.07 (488), 0.00 (6), 0.00 (46)
r-NP-IS	−0.061	(−0.137, 0.015)	0-2-2	−0.15 (456), −0.09 (46), 0.00 (488)

3.3 Icelandic Retrieval

For Icelandic, we used English stopwords and English stemming. We review some topics to see what can be learned about Icelandic retrieval.

WC0488: Table 9 shows that the only topic on which English stemming made a difference was topic 488 (framboð ferskvatns í evrópu (Fresh water supplies in europa)). The desired page's rank fell from 1 to 2 with English stemming because it matched the word 'Ferskvatn' which was not in the desired page (the English lexical stemmer was augmented with a stem guesser for unrecognized words). A variant in the desired page, 'ferskvatni', was not matched by English stemming. It appears that 'í' is a potential Icelandic stopword ('i' actually was not in our English list though arguably should be). This topic also shows that Icelandic uses the small letter Eth (ð). SearchServer case normalizes ð to the capital letter Eth (Ð).

WC0456: In topic 456 (upplýsingar um europol (europol factsheet)), English stemming missed apparent variants to the query word 'upplýsingar' such as 'Upplýsingasíða' and 'upplýsingamál'. 'um' appears to be a potential Icelandic stopword.

WC0243: In (home page) topic 243 (umhverfisstofnun evrópu (european environment agency)), we noticed that some web pages used entity references such as '&eth;' and '&thorn;' and '&yacute;' which our cTREC text reader did not normalize to the corresponding character, possibly impairing results for some queries.

We were disappointed that the Icelandic thorn (lowercase þ or uppercase Þ) was not used in any of the topic words. But overall, even with just 5 topics in the test set, we have learned at least that an Icelandic stemmer would potentially be helpful for Icelandic retrieval.

Table 10: Mean Scores of WebCLEF Runs on English Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dplD-NP-EN-other	0.761	25/56 (45%)	38/56 (68%)	44/56 (79%)	0.570
dpD-NP-EN-other	0.737	26/56 (46%)	39/56 (70%)	42/56 (75%)	0.579
dpD0-NP-EN-other	0.737	26/56 (46%)	39/56 (70%)	42/56 (75%)	0.579
dp-NP-EN-other	0.690	24/56 (43%)	37/56 (66%)	38/56 (68%)	0.541
p-NP-EN-other	0.690	23/56 (41%)	37/56 (66%)	38/56 (68%)	0.535
rdp-NP-EN-other	0.684	24/56 (43%)	38/56 (68%)	38/56 (68%)	0.531
none-NP-EN-other	0.678	23/56 (41%)	37/56 (66%)	38/56 (68%)	0.525
dplD-HP-EN-other	0.652	14/35 (40%)	22/35 (63%)	24/35 (69%)	0.499
dpD-HP-EN-other	0.572	12/35 (34%)	18/35 (51%)	21/35 (60%)	0.432
dpD0-HP-EN-other	0.572	12/35 (34%)	18/35 (51%)	21/35 (60%)	0.432
dp-HP-EN-other	0.544	12/35 (34%)	18/35 (51%)	19/35 (54%)	0.426
p-HP-EN-other	0.531	12/35 (34%)	17/35 (49%)	18/35 (51%)	0.413
rdp-HP-EN-other	0.472	11/35 (31%)	15/35 (43%)	16/35 (46%)	0.380
none-HP-EN-other	0.399	9/35 (26%)	12/35 (34%)	14/35 (40%)	0.302
dplD-NP-EN-hum	0.956	11/15 (73%)	14/15 (93%)	15/15 (100%)	0.832
dpD-NP-EN-hum	0.956	11/15 (73%)	14/15 (93%)	15/15 (100%)	0.832
dpD0-NP-EN-hum	0.956	11/15 (73%)	14/15 (93%)	15/15 (100%)	0.832
dp-NP-EN-hum	0.836	8/15 (53%)	12/15 (80%)	13/15 (87%)	0.651
rdp-NP-EN-hum	0.833	8/15 (53%)	12/15 (80%)	13/15 (87%)	0.651
p-NP-EN-hum	0.832	9/15 (60%)	12/15 (80%)	13/15 (87%)	0.682
none-NP-EN-hum	0.803	9/15 (60%)	12/15 (80%)	12/15 (80%)	0.686
rdp-HP-EN-hum	0.538	6/15 (40%)	8/15 (53%)	8/15 (53%)	0.461
dplD-HP-EN-hum	0.521	7/15 (47%)	7/15 (47%)	7/15 (47%)	0.480
dpD-HP-EN-hum	0.516	6/15 (40%)	7/15 (47%)	7/15 (47%)	0.432
dpD0-HP-EN-hum	0.516	6/15 (40%)	7/15 (47%)	7/15 (47%)	0.432
dp-HP-EN-hum	0.490	6/15 (40%)	7/15 (47%)	7/15 (47%)	0.427
p-HP-EN-hum	0.464	6/15 (40%)	7/15 (47%)	7/15 (47%)	0.418
none-HP-EN-hum	0.410	4/15 (27%)	6/15 (40%)	6/15 (40%)	0.323

3.4 English Topic Contributions

WebCLEF participants were requested to contribute at least 30 known-item topics. Each topic consisted of a query, the correct answer page in EuroGOV, and a list of duplicate and translated pages in EuroGOV. We contributed 30 English topics. Tables 10 and 11 separate the results for our topics from the other English topics. Based on the scores, it appears that our named page topics may have been easier than the others, but our home page topics may have been harder.

To create a topic, we typically started by randomly selecting an English language page from the EuroGOV collection. (The organizers had provided a languages.tar.gz file which listed the languages detected in each document; we reduced this file to the 252,574 pages labelled just as 'english', then randomly selected pages from this list.) We alternated between creating named page queries and home page queries.

If we wanted a named page query, we tried to understand the random page well enough to create an unambiguous query for it. Sometimes we rejected a page for being too obscure, and tried browsing to a related page for which a clearer query could be made. (Browsing was done on the live web; then we would find the new page in EuroGOV by extracting a phrase and searching EuroGOV with SearchServer.) If browsing was not fruitful, we started over with a new random page. Sometimes we started over because the area we were browsing looked too similar to an area for which we had already made a query.

Table 11: Impact of Web Techniques on First Relevant Score, English Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
o-NP-EN-oth	0.047	(0.008, 0.087)	12-0-44	0.87 (292), 0.54 (384), 0.00 (532)
l-NP-EN-oth	0.023	(−0.016, 0.062)	10-8-38	0.74 (2), 0.45 (479), −0.30 (165)
p-NP-EN-oth	0.011	(−0.030, 0.052)	15-5-36	−0.83 (292), −0.32 (38), 0.50 (76)
s-NP-EN-oth	0.000	n/a	0-0-56	0.00 (333), 0.00 (34), 0.00 (532)
d-NP-EN-oth	0.000	(−0.010, 0.011)	5-7-44	0.15 (314), 0.11 (38), −0.14 (423)
r-NP-EN-oth	−0.006	(−0.036, 0.024)	4-12-40	0.68 (418), −0.19 (91), −0.27 (88)
p-HP-EN-oth	0.133	(0.042, 0.224)	15-1-19	1.00 (101), 0.98 (313), −0.07 (436)
l-HP-EN-oth	0.080	(0.010, 0.150)	9-2-24	0.80 (1), 0.75 (275), −0.13 (246)
o-HP-EN-oth	0.028	(0.000, 0.056)	8-0-27	0.38 (287), 0.23 (85), 0.00 (539)
d-HP-EN-oth	0.012	(−0.003, 0.028)	6-5-24	0.18 (190), 0.13 (100), −0.07 (436)
s-HP-EN-oth	0.000	n/a	0-0-35	0.00 (275), 0.00 (85), 0.00 (539)
r-HP-EN-oth	−0.072	(−0.148, 0.004)	3-10-22	−0.93 (246), −0.86 (190), 0.17 (335)
o-NP-EN-hum	0.120	(−0.028, 0.268)	3-0-12	1.00 (285), 0.54 (129), 0.00 (538)
p-NP-EN-hum	0.029	(−0.030, 0.088)	2-1-12	0.43 (129), 0.08 (325), −0.07 (295)
d-NP-EN-hum	0.003	(−0.014, 0.021)	2-1-12	0.09 (325), 0.03 (129), −0.07 (139)
l-NP-EN-hum	0.000	(−0.015, 0.015)	1-1-13	0.07 (139), 0.00 (129), −0.07 (513)
s-NP-EN-hum	0.000	n/a	0-0-15	0.00 (295), 0.00 (94), 0.00 (538)
r-NP-EN-hum	−0.003	(−0.024, 0.018)	1-2-12	0.10 (325), −0.05 (513), −0.10 (129)
p-HP-EN-hum	0.054	(0.000, 0.109)	5-0-10	0.37 (408), 0.21 (167), 0.00 (507)
r-HP-EN-hum	0.048	(−0.058, 0.154)	3-3-9	0.69 (476), 0.21 (408), −0.23 (345)
o-HP-EN-hum	0.026	(−0.027, 0.080)	1-0-14	0.40 (507), 0.00 (167), 0.00 (207)
d-HP-EN-hum	0.025	(−0.003, 0.053)	4-1-10	0.17 (476), 0.12 (345), −0.01 (399)
l-HP-EN-hum	0.005	(−0.030, 0.040)	2-3-10	0.21 (408), 0.04 (476), −0.13 (507)
s-HP-EN-hum	0.000	n/a	0-0-15	0.00 (207), 0.00 (141), 0.00 (507)

For home page queries, usually the random start page was not a home page, so we would typically try to browse to the closest home page for that page (again, typically on the live web, by following links or truncating the url).

To find duplicates, typically we extracted a phrase from the document and used SearchServer to find other pages with that phrase, then checked those pages to confirm they were duplicates. If a page had more duplicates than we were willing to record, we started over with a new page.

To find translations, typically we browsed the live web for links to translated pages, then used SearchServer to find them in EuroGOV. Finding translations took a lot of detective work. Sometimes the url was the same except for a language tag, making it easy to find the translations with SearchServer. Sometimes sites had direct links to the translations, which was also easy. But sometimes sites just had links to the top-level page for each language, so we would see how to browse down for English, and then try to do analogous browsing for the translation language, grasping for clues such as possible word translations or similar pictures, to get to the proper translated page. It's quite possible we missed some translations.

For the query itself, we tried to make it as realistic as possible (e.g. short and general) but also unambiguous. This could depend on what other pages were available; e.g. for a biography of Giuseppe Medici, it was enough just to specify 'Giuseppe Medici' as the query because there were no other (English) pages focused on that person. Usually we tried candidate queries with the organizer-provided engines or a web search engine to see if there might be other valid interpretations of the query we hadn't expected, so that we could adjust the query accordingly.

It seemed that a lot of times, our query ended up being fairly similar to the document title. Table 11 shows that in the 'p' experiment (which isolates giving more weight to the title and other meta properties), our queries did tend to be helped by weighting the title more. But the other groups' English queries actually benefited even more often from this weighting.

Table 12: Mean Scores of WebCLEF Runs on Spanish Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dpD-NP-ES	0.758	32/67 (48%)	47/67 (70%)	53/67 (79%)	0.595
dplD-NP-ES	0.720	27/67 (40%)	42/67 (63%)	52/67 (78%)	0.529
dpD0-NP-ES	0.670	26/67 (39%)	42/67 (63%)	47/67 (70%)	0.497
p-NP-ES	0.650	25/67 (37%)	41/67 (61%)	46/67 (69%)	0.478
dp-NP-ES	0.648	26/67 (39%)	40/67 (60%)	44/67 (66%)	0.486
rdp-NP-ES	0.639	25/67 (37%)	39/67 (58%)	43/67 (64%)	0.475
none-NP-ES	0.624	21/67 (31%)	39/67 (58%)	42/67 (63%)	0.433
dplD-HP-ES	0.446	15/67 (22%)	26/67 (39%)	31/67 (46%)	0.297
rdp-HP-ES	0.437	15/67 (22%)	26/67 (39%)	29/67 (43%)	0.307
dpD-HP-ES	0.388	11/67 (16%)	21/67 (31%)	27/67 (40%)	0.240
dpD0-HP-ES	0.369	11/67 (16%)	20/67 (30%)	25/67 (37%)	0.235
dp-HP-ES	0.364	11/67 (16%)	20/67 (30%)	24/67 (36%)	0.234
p-HP-ES	0.325	9/67 (13%)	19/67 (28%)	23/67 (34%)	0.201
none-HP-ES	0.279	5/67 (7%)	13/67 (19%)	18/67 (27%)	0.142

Table 13: Impact of Web Techniques on First Relevant Score, Spanish Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
s-NP-ES	0.087	(0.042, 0.133)	20-4-43	0.87 (116), 0.58 (344), −0.10 (45)
p-NP-ES	0.026	(0.000, 0.052)	16-10-41	0.59 (118), 0.30 (200), −0.26 (449)
o-NP-ES	0.022	(0.006, 0.038)	13-0-54	0.34 (489), 0.25 (502), 0.00 (544)
d-NP-ES	−0.002	(−0.011, 0.007)	7-10-50	0.12 (309), −0.10 (84), −0.11 (502)
r-NP-ES	−0.010	(−0.021, 0.002)	2-14-51	−0.20 (98), −0.18 (309), 0.19 (45)
l-NP-ES	−0.037	(−0.075, 0.000)	7-21-39	−0.78 (157), −0.39 (330), 0.34 (483)
r-HP-ES	0.073	(0.021, 0.126)	16-7-44	0.92 (32), 0.92 (542), −0.12 (13)
l-HP-ES	0.058	(0.000, 0.115)	13-14-40	0.96 (123), 0.93 (124), −0.56 (397)
p-HP-ES	0.045	(−0.003, 0.094)	15-8-44	0.88 (393), 0.78 (468), −0.54 (473)
d-HP-ES	0.039	(0.012, 0.066)	16-3-48	0.50 (414), 0.43 (220), −0.07 (299)
s-HP-ES	0.019	(0.002, 0.037)	15-4-48	0.31 (522), 0.27 (543), −0.18 (467)
o-HP-ES	0.005	(−0.001, 0.010)	6-0-61	0.13 (130), 0.08 (154), 0.00 (543)

3.5 Other Languages

Unfortunately, we have run out of time to walk through topics for more languages. But for future reference, we list the per-topic tables for the remaining languages (in descending order by number of topics).

Table 14: Mean Scores of WebCLEF Runs on Dutch Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dpD-NP-NL	0.958	26/34 (76%)	33/34 (97%)	33/34 (97%)	0.860
p-NP-NL	0.952	27/34 (79%)	33/34 (97%)	33/34 (97%)	0.864
dpD0-NP-NL	0.951	27/34 (79%)	33/34 (97%)	33/34 (97%)	0.865
dp-NP-NL	0.946	26/34 (76%)	33/34 (97%)	33/34 (97%)	0.845
none-NP-NL	0.936	26/34 (76%)	31/34 (91%)	33/34 (97%)	0.833
dplD-NP-NL	0.918	24/34 (71%)	30/34 (88%)	31/34 (91%)	0.791
rdp-NP-NL	0.903	25/34 (74%)	30/34 (88%)	32/34 (94%)	0.804
dpD-HP-NL	0.723	9/25 (36%)	16/25 (64%)	19/25 (76%)	0.496
dplD-HP-NL	0.688	10/25 (40%)	15/25 (60%)	18/25 (72%)	0.488
dpD0-HP-NL	0.649	8/25 (32%)	15/25 (60%)	16/25 (64%)	0.445
dp-HP-NL	0.649	8/25 (32%)	15/25 (60%)	16/25 (64%)	0.445
p-HP-NL	0.617	8/25 (32%)	13/25 (52%)	18/25 (72%)	0.419
rdp-HP-NL	0.607	7/25 (28%)	14/25 (56%)	16/25 (64%)	0.390
none-HP-NL	0.571	5/25 (20%)	13/25 (52%)	15/25 (60%)	0.348

Table 15: Impact of Web Techniques on First Relevant Score, Dutch Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
p-NP-NL	0.016	(−0.005, 0.037)	3-0-31	0.30 (296), 0.18 (516), 0.00 (547)
s-NP-NL	0.006	(−0.006, 0.018)	3-1-30	0.15 (547), 0.07 (296), −0.07 (308)
o-NP-NL	0.005	(−0.001, 0.012)	3-0-31	0.07 (269), 0.07 (438), 0.00 (386)
d-NP-NL	−0.006	(−0.016, 0.005)	1-3-30	−0.12 (516), −0.07 (3), 0.07 (269)
l-NP-NL	−0.040	(−0.090, 0.011)	2-5-27	−0.60 (509), −0.50 (338), 0.13 (547)
r-NP-NL	−0.043	(−0.103, 0.017)	2-3-29	−0.87 (469), −0.46 (528), 0.07 (269)
s-HP-NL	0.075	(−0.033, 0.183)	6-4-15	0.91 (39), 0.68 (486), −0.30 (506)
p-HP-NL	0.046	(−0.017, 0.108)	7-3-15	0.50 (90), 0.41 (75), −0.19 (140)
d-HP-NL	0.032	(−0.028, 0.092)	7-5-13	−0.40 (290), 0.32 (358), 0.36 (535)
o-HP-NL	0.000	n/a	0-0-25	0.00 (221), 0.00 (26), 0.00 (546)
l-HP-NL	−0.035	(−0.097, 0.026)	2-5-18	−0.68 (324), −0.21 (140), 0.17 (517)
r-HP-NL	−0.041	(−0.155, 0.072)	2-8-15	0.84 (39), 0.52 (21), −0.59 (67)

Table 16: Mean Scores of WebCLEF Runs on Portuguese Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dplD-NP-PT	0.579	5/30 (17%)	16/30 (53%)	18/30 (60%)	0.325
dpD0-NP-PT	0.534	4/30 (13%)	13/30 (43%)	20/30 (67%)	0.276
dpD-NP-PT	0.551	6/30 (20%)	13/30 (43%)	18/30 (60%)	0.328
rdp-NP-PT	0.532	4/30 (13%)	11/30 (37%)	16/30 (53%)	0.275
dp-NP-PT	0.516	4/30 (13%)	12/30 (40%)	18/30 (60%)	0.264
p-NP-PT	0.511	4/30 (13%)	12/30 (40%)	18/30 (60%)	0.270
none-NP-PT	0.469	2/30 (7%)	11/30 (37%)	17/30 (57%)	0.219
rdp-HP-PT	0.665	14/29 (48%)	18/29 (62%)	19/29 (66%)	0.546
dpD-HP-PT	0.628	13/29 (45%)	16/29 (55%)	18/29 (62%)	0.507
dplD-HP-PT	0.621	14/29 (48%)	16/29 (55%)	17/29 (59%)	0.522
dpD0-HP-PT	0.545	11/29 (38%)	14/29 (48%)	17/29 (59%)	0.438
dp-HP-PT	0.544	11/29 (38%)	14/29 (48%)	17/29 (59%)	0.438
p-HP-PT	0.435	8/29 (28%)	10/29 (34%)	13/29 (45%)	0.324
none-HP-PT	0.263	3/29 (10%)	6/29 (21%)	7/29 (24%)	0.148

Table 17: Impact of Web Techniques on First Relevant Score, Portuguese Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
p-NP-PT	0.041	(−0.018, 0.101)	8-4-18	0.59 (415), 0.58 (303), −0.19 (4)
l-NP-PT	0.028	(−0.037, 0.094)	5-9-16	0.86 (248), 0.30 (226), −0.17 (4)
o-NP-PT	0.018	(−0.009, 0.046)	3-0-27	0.40 (377), 0.08 (216), 0.00 (529)
s-NP-PT	0.017	(−0.007, 0.041)	9-5-16	0.19 (415), 0.14 (215), −0.10 (4)
r-NP-PT	0.016	(−0.017, 0.049)	5-6-19	0.34 (69), 0.23 (377), −0.10 (4)
d-NP-PT	0.005	(−0.024, 0.034)	2-4-24	0.40 (377), −0.07 (215), −0.08 (303)
p-HP-PT	0.172	(0.081, 0.262)	15-0-14	0.71 (390), 0.71 (163), 0.00 (260)
r-HP-PT	0.121	(0.034, 0.207)	13-2-14	0.86 (362), 0.72 (96), −0.07 (52)
d-HP-PT	0.110	(0.028, 0.192)	11-0-18	0.91 (52), 0.75 (381), 0.00 (260)
s-HP-PT	0.083	(0.031, 0.135)	10-0-19	0.43 (362), 0.37 (96), 0.00 (164)
o-HP-PT	0.000	(−0.001, 0.001)	1-0-28	0.01 (326), 0.00 (33), 0.00 (545)
l-HP-PT	−0.006	(−0.018, 0.005)	1-4-24	−0.10 (526), −0.08 (382), 0.07 (164)

Table 18: Mean Scores of WebCLEF Runs on German Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dplD-NP-DE	0.706	16/34 (47%)	24/34 (71%)	25/34 (74%)	0.556
dpD-NP-DE	0.628	13/34 (38%)	21/34 (62%)	22/34 (65%)	0.484
rdp-NP-DE	0.606	15/34 (44%)	19/34 (56%)	22/34 (65%)	0.495
dpD0-NP-DE	0.602	14/34 (41%)	19/34 (56%)	21/34 (62%)	0.480
dp-NP-DE	0.595	14/34 (41%)	19/34 (56%)	21/34 (62%)	0.478
p-NP-DE	0.591	14/34 (41%)	19/34 (56%)	21/34 (62%)	0.479
none-NP-DE	0.589	12/34 (35%)	19/34 (56%)	20/34 (59%)	0.444
dpD-HP-DE	0.526	4/23 (17%)	11/23 (48%)	12/23 (52%)	0.306
dpD0-HP-DE	0.518	3/23 (13%)	10/23 (43%)	13/23 (57%)	0.259
dp-HP-DE	0.512	3/23 (13%)	10/23 (43%)	13/23 (57%)	0.257
dplD-HP-DE	0.472	4/23 (17%)	10/23 (43%)	11/23 (48%)	0.295
rdp-HP-DE	0.466	5/23 (22%)	9/23 (39%)	12/23 (52%)	0.296
p-HP-DE	0.451	2/23 (9%)	8/23 (35%)	11/23 (48%)	0.219
none-HP-DE	0.385	2/23 (9%)	7/23 (30%)	10/23 (43%)	0.189

Table 19: Impact of Web Techniques on First Relevant Score, German Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
l-NP-DE	0.078	(−0.005, 0.161)	8-5-21	0.95 (402), 0.79 (212), −0.13 (288)
s-NP-DE	0.026	(−0.021, 0.074)	6-5-23	0.76 (477), 0.25 (351), −0.07 (339)
r-NP-DE	0.011	(−0.019, 0.041)	3-4-27	0.46 (477), 0.07 (347), −0.14 (197)
o-NP-DE	0.007	(−0.006, 0.020)	3-0-31	0.21 (197), 0.02 (316), 0.00 (536)
d-NP-DE	0.004	(−0.010, 0.017)	3-5-26	0.16 (197), 0.09 (536), −0.07 (95)
p-NP-DE	0.002	(−0.047, 0.051)	7-5-22	−0.64 (477), 0.25 (95), 0.31 (351)
p-HP-DE	0.066	(−0.018, 0.149)	7-4-12	0.86 (300), 0.28 (241), −0.11 (10)
d-HP-DE	0.062	(−0.034, 0.157)	5-2-16	1.00 (453), 0.35 (47), −0.13 (20)
s-HP-DE	0.007	(−0.084, 0.099)	7-3-13	−0.79 (20), 0.32 (412), 0.42 (47)
o-HP-DE	0.006	(0.000, 0.013)	4-0-19	0.05 (433), 0.04 (412), 0.00 (453)
r-HP-DE	−0.047	(−0.126, 0.032)	2-9-12	−0.72 (20), −0.29 (236), 0.42 (47)
l-HP-DE	−0.054	(−0.114, 0.007)	4-10-9	−0.47 (133), −0.40 (214), 0.19 (396)

Table 20: Mean Scores of WebCLEF Runs on Hungarian Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
p-NP-HU	0.766	12/19 (63%)	14/19 (74%)	15/19 (79%)	0.665
dpD0-NP-HU	0.763	12/19 (63%)	14/19 (74%)	15/19 (79%)	0.664
rdp-NP-HU	0.763	12/19 (63%)	14/19 (74%)	15/19 (79%)	0.664
dp-NP-HU	0.763	12/19 (63%)	14/19 (74%)	15/19 (79%)	0.664
none-NP-HU	0.763	9/19 (47%)	15/19 (79%)	15/19 (79%)	0.595
dpD-NP-HU	0.706	9/19 (47%)	12/19 (63%)	14/19 (74%)	0.559
dplD-NP-HU	0.656	9/19 (47%)	12/19 (63%)	13/19 (68%)	0.533
dpD0-HP-HU	0.579	3/16 (19%)	9/16 (56%)	10/16 (63%)	0.326
dpD-HP-HU	0.575	4/16 (25%)	8/16 (50%)	9/16 (56%)	0.362
dp-HP-HU	0.569	3/16 (19%)	8/16 (50%)	10/16 (63%)	0.322
dplD-HP-HU	0.553	4/16 (25%)	7/16 (44%)	8/16 (50%)	0.352
rdp-HP-HU	0.543	2/16 (13%)	5/16 (31%)	10/16 (63%)	0.265
p-HP-HU	0.433	3/16 (19%)	4/16 (25%)	7/16 (44%)	0.262
none-HP-HU	0.415	4/16 (25%)	4/16 (25%)	5/16 (31%)	0.288

Table 21: Impact of Web Techniques on First Relevant Score, Hungarian Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
p-NP-HU	0.002	(−0.031, 0.036)	3-1-15	−0.25 (448), 0.07 (283), 0.14 (527)
o-NP-HU	0.000	n/a	0-0-19	0.00 (283), 0.00 (102), 0.00 (527)
r-NP-HU	0.000	n/a	0-0-19	0.00 (283), 0.00 (102), 0.00 (527)
d-NP-HU	−0.003	(−0.008, 0.003)	0-1-18	−0.05 (448), 0.00 (527), 0.00 (283)
l-NP-HU	−0.050	(−0.119, 0.019)	2-5-12	−0.59 (225), −0.26 (110), 0.07 (463)
s-NP-HU	−0.057	(−0.150, 0.036)	0-4-15	−0.88 (527), −0.07 (24), 0.00 (283)
d-HP-HU	0.136	(−0.006, 0.279)	7-3-6	0.78 (435), 0.74 (43), −0.14 (245)
p-HP-HU	0.017	(−0.135, 0.169)	5-2-9	−0.92 (435), 0.45 (346), 0.54 (148)
o-HP-HU	0.009	(−0.010, 0.029)	1-0-15	0.15 (51), 0.00 (49), 0.00 (510)
s-HP-HU	−0.003	(−0.116, 0.110)	6-4-6	−0.76 (346), 0.19 (494), 0.26 (43)
l-HP-HU	−0.022	(−0.101, 0.057)	3-4-9	−0.50 (51), −0.14 (298), 0.27 (9)
r-HP-HU	−0.026	(−0.236, 0.183)	7-7-2	−1.00 (148), −0.66 (346), 0.64 (494)

Table 22: Mean Scores of WebCLEF Runs on Russian Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dplD-NP-RU	0.401	5/15 (33%)	6/15 (40%)	6/15 (40%)	0.354
dpD-NP-RU	0.392	4/15 (27%)	6/15 (40%)	6/15 (40%)	0.335
dpD0-NP-RU	0.381	4/15 (27%)	6/15 (40%)	6/15 (40%)	0.317
p-NP-RU	0.372	3/15 (20%)	6/15 (40%)	6/15 (40%)	0.272
dp-NP-RU	0.368	3/15 (20%)	6/15 (40%)	6/15 (40%)	0.267
rdp-NP-RU	0.362	4/15 (27%)	5/15 (33%)	6/15 (40%)	0.293
none-NP-RU	0.357	3/15 (20%)	5/15 (33%)	6/15 (40%)	0.260
rdp-HP-RU	0.359	0/15 (0%)	6/15 (40%)	6/15 (40%)	0.134
dpD-HP-RU	0.355	1/15 (7%)	5/15 (33%)	5/15 (33%)	0.163
dpD0-HP-RU	0.300	0/15 (0%)	3/15 (20%)	5/15 (33%)	0.084
dp-HP-RU	0.300	0/15 (0%)	3/15 (20%)	5/15 (33%)	0.084
dplD-HP-RU	0.282	2/15 (13%)	4/15 (27%)	5/15 (33%)	0.174
p-HP-RU	0.249	0/15 (0%)	2/15 (13%)	4/15 (27%)	0.069
none-HP-RU	0.174	0/15 (0%)	2/15 (13%)	3/15 (20%)	0.044

Table 23: Impact of Web Techniques on First Relevant Score, Russian Queries

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
p-NP-RU	0.015	(−0.016, 0.046)	1-0-14	0.23 (457), 0.00 (63), 0.00 (540)
o-NP-RU	0.014	(−0.014, 0.042)	1-0-14	0.21 (359), 0.00 (63), 0.00 (540)
s-NP-RU	0.011	(−0.007, 0.029)	2-0-13	0.13 (457), 0.03 (63), 0.00 (540)
l-NP-RU	0.008	(−0.020, 0.036)	3-1-11	−0.13 (457), 0.07 (83), 0.12 (63)
d-NP-RU	−0.004	(−0.013, 0.005)	0-1-14	−0.06 (457), 0.00 (540), 0.00 (263)
r-NP-RU	−0.006	(−0.031, 0.019)	1-1-13	−0.16 (457), 0.00 (63), 0.07 (83)
p-HP-RU	0.075	(−0.003, 0.153)	5-1-9	0.39 (71), 0.39 (136), −0.05 (466)
r-HP-RU	0.059	(−0.073, 0.191)	3-2-10	0.79 (181), 0.32 (520), −0.26 (136)
s-HP-RU	0.055	(−0.017, 0.127)	4-1-10	0.46 (520), 0.23 (22), −0.11 (17)
d-HP-RU	0.051	(−0.021, 0.123)	5-0-10	0.54 (520), 0.11 (466), 0.00 (240)
o-HP-RU	0.000	n/a	0-0-15	0.00 (240), 0.00 (22), 0.00 (520)
l-HP-RU	−0.073	(−0.153, 0.007)	2-6-7	−0.46 (136), −0.27 (240), 0.14 (22)

Table 24: Mean Scores of WebCLEF Runs on French Queries

Run	FRS	Success@1	Success@5	Success@10	MRR
dplD-NP-FR	1.000	1/1 (100%)	1/1 (100%)	1/1 (100%)	1.000
dpD-NP-FR	1.000	1/1 (100%)	1/1 (100%)	1/1 (100%)	1.000
dpD0-NP-FR	1.000	1/1 (100%)	1/1 (100%)	1/1 (100%)	1.000
rdp-NP-FR	1.000	1/1 (100%)	1/1 (100%)	1/1 (100%)	1.000
dp-NP-FR	1.000	1/1 (100%)	1/1 (100%)	1/1 (100%)	1.000
p-NP-FR	1.000	1/1 (100%)	1/1 (100%)	1/1 (100%)	1.000
none-NP-FR	1.000	1/1 (100%)	1/1 (100%)	1/1 (100%)	1.000

References

[1] AltaVista's Babel Fish Translation Service. http://babelfish.altavista.com/tr

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

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

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

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

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

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

[8] Börkur Sigurbjörnsson, Jaap Kamps and Maarten de Rijke. EuroGOV: Engineering a Multilingual Web Corpus. To appear in *Working Notes for the CLEF 2005 Workshop.*

[9] Börkur Sigurbjörnsson, Jaap Kamps and Maarten de Rijke. Overview of WebCLEF 2005. To appear in *Working Notes for the CLEF 2005 Workshop.*

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

[11] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer^{TM} at CLEF 2005. To appear in *Working Notes for the CLEF 2005 Workshop.*

[12] Stephen Tomlinson. Experiments in Named Page Finding and Arabic Retrieval with Hummingbird SearchServer^{TM} at TREC 2002. *Proceedings of TREC 2002.*

[13] Stephen Tomlinson. Robust, Web and Genomic Retrieval with Hummingbird SearchServer^{TM} at TREC 2003. *Proceedings of TREC 2003.*

[14] Stephen Tomlinson. Robust, Web and Terabyte Retrieval with Hummingbird SearchServer^{TM} at TREC 2004. *Proceedings of TREC 2004.*

[15] Thijs Westerveld, Wessel Kraaij and Djoerd Hiemstra. Retrieving Web Pages using Content, Links, URLs and Anchors. *Proceedings of TREC 2001.*