Finnish, Portuguese and Russian Retrieval with Hummingbird SearchServer™ at CLEF 2004

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

August 16, 2004

Abstract

Hummingbird participated in the Finnish, Portuguese, Russian and French monolingual information retrieval tasks of the Cross-Language Evaluation Forum (CLEF) 2004: for the natural language queries, find all the relevant documents (with high precision) in the CLEF 2004 document sets. SearchServer's experimental lexical stemmers significantly increased mean average precision for each of the 4 languages. For Finnish, mean average precision was significantly higher with SearchServer's experimental decompounding option enabled. For each language, the submitted SearchServer run returned a relevant document in the first row for more than half of the short (Title-only) queries. At least one relevant document was returned in the first ten rows for 75-90% of the short queries.

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.

SearchServer works in Unicode internally [3] and supports most of the world's major character sets and languages. The major conferences in text retrieval evaluation (CLEF [1], NTCIR [4] and TREC [8]) have provided opportunities to objectively evaluate SearchServer's support for more than a dozen languages.

This (draft) paper describes experimental work with SearchServer for the task of finding relevant documents for natural language queries in 4 European languages (Finnish, Portuguese, Russian and French) using the CLEF 2004 test collections. Portuguese is new to CLEF this year, and the experimental SearchServer version has some enhancements which substantially affect Finnish and Russian, so we focus on these 3 languages.

2 Methodology

2.1 Data

The CLEF 2004 document sets consisted of tagged (SGML-formatted) news articles (mostly from 1995) in 4 different languages: Finnish, Portuguese, Russian and French. Table 1 gives the sizes.

The CLEF organizers created 50 natural language "topics" (numbered 201-250) and translated them into many languages. Each topic contained a "Title" (subject of the topic), "Description"

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

Table 1: Sizes of CLEF 2004 Test Collections

Language	Text Size (uncompressed)	Documents	Topics	Rel/Topic
French	255,334,872 bytes (244 MB)	90,261	49	19
Portuguese	185,739,565 bytes (177 MB)	55,070	46	15
Finnish	143,902,109 bytes (137 MB)	55,344	45	9
Russian	68,802,653 bytes (66 MB)	16,716	34	4

(a one-sentence specification of the information need) and "Narrative" (more detailed guidelines for what a relevant document should or should not contain). The participants were asked to use the Title and Description fields for at least one automatic submission per task this year to facilitate comparison of results. Some topics were discarded for some languages because no relevant documents existed for them. Table 1 gives the final number of topics for each language and their average number of relevant documents. For more information on the CLEF test collections, see the CLEF web site [1].

2.2 Indexing

Our indexing approach was the mostly the same as last year [10]. Accents were not indexed except for the combining breve in Russian. The apostrophe was treated as a word separator for the 4 investigated languages. Our custom text reader, cTREC, was updated to maintain support for the CLEF guidelines of only indexing specifically tagged fields (the new Portuguese collection necessitated a minor update).

Some stop words were excluded from indexing (e.g. "the", "by" and "of" in English). For these experiments, our stop word lists for Portuguese and Russian were based on the Porter lists [5], and this year we based on our Finnish list on Savoy's [7]. We used our own list for French.

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).

For many languages (including the 4 European languages investigated in CLEF 2004), SearchServer includes the option of finding inflections based on lexical stemming (i.e. stemming based on a dictionary or lexicon for the language). For example, in English, "baby", "babied", "babies", "baby's" and "babying" all have "baby" as a stem. Specifying an inflected search for any of these terms will match all of the others. The lexical stemming of the post-5.x experimental development version of SearchServer used for the experiments in this paper was based on internal stemming component 3.6.3.4 for the submitted runs and 3.7.0.15 for the diagnostic runs. We treat each linguistic component as a black box in this paper.

SearchServer typically does "inflectional" stemming which generally retains the part of speech (e.g. a plural of a noun is typically stemmed to the singular form). It typically does not do "derivational" stemming which would often change the part of speech or the meaning more substantially (e.g. "performer" is not stemmed to "perform").

SearchServer's lexical stemming includes compound-splitting (decompounding) for compound words in Finnish (and also some other languages not investigated this year, such as German, Dutch and Swedish). For example, in German, "babykost" (baby food) has "baby" and "kost" as stems.

SearchServer's lexical stemming also supports some spelling variations. In English, British and American spellings have the same stems, e.g. "labour" stems to "labor", "hospitalisation" stems to "hospitalization" and "plough" stems to "plow".

Lexical stemmers can produce more than one stem, even for non-compound words. For example, in English, "axes" has both "axe" and "axis" as stems (different meanings), and in French, "important" has both "important" (adjective) and "importer" (verb) as stems (different parts of speech). SearchServer records all the stem mappings at index-time to support maximum recall and does so in a way to allow searching to weight some inflections higher than others.

2.3 Searching

Unlike previous years, this year we experimented with SearchServer's CONTAINS predicate (instead of the IS_ABOUT predicate) though it should not make a difference to the ranking. Our test application specified SearchSQL to perform a boolean-OR of the query words. For example, for Russian topic 250 whose Title was "Бешенство у людей" (Rabies in Humans), a corresponding SearchSQL query would be:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM CLEF04RU
WHERE FT_TEXT CONTAINS 'Бешенство'|'у'|'людей'
ORDER BY REL DESC;
```

(Note that "y" is a stopword for Russian so its inclusion in the query won't actually add any matches.)

Most aspects of SearchServer's relevance value calculation are the same as described last year [10]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [6] and dampens the inverse document frequency using an approximation of the logarithm. These calculations are based on the stems of the terms when doing morphological searching (i.e. when SET TERM_GENERATOR 'word!ftelp/inflect' was previously specified).

An experimental new default is that SearchServer only includes morphological matches from compound words if all of its stems (from a particular stemming interpretation) are in the same or consecutive words. For example, in German, a morphological search for the compound "babykost" (baby food) will no longer match "baby" or "kost" by themselves, but it will match "babykost" and "baby kost" (and if SET PHRASE_DISTANCE 1 is specified, it will also match the hyphenated "baby-kost"). Words (and compounds) still match inside compounds (and larger compounds), e.g. a search for "kost" still matches "babykost". To restore the old behaviour of matching if just one stem is in common, one can specify the /decompound option (e.g. SET TERM_GENERATOR 'word!ftelp/inflect/decompound'). See Section 3.3.1 for several more decompounding examples.

This year's experimental SearchServer version contains an enhancement for handling multiple stemming interpretations. For each document, only the interpretation that produces the highest score for the document is used in the relevance calculation (but all interpretations are still used for matching and search term highlighting). Sometimes this enhancement causes the original query form of the word to get more weight than some of its inflections (and it never gets less weight). This approach overcomes the previous issue of terms with multiple stemming interpretations being over-weighted; it used to be better for CLEF experiments to workaround by using the /single or /noalt options, but Section 3.5 verifies that this is no longer the case.

SearchServer's RELEVANCE_METHOD setting can be used to optionally square the importance of the inverse document frequency (by choosing a RELEVANCE_METHOD of '2:4' instead of '2:3'). The importance of document length to the ranking is controlled by SearchServer's RELEVANCE_DLEN_IMP setting (scale of 0 to 1000). For all experiments in this paper, RELEVANCE_METHOD was set to '2:3' and RELEVANCE_DLEN_IMP was set to 750.

2.4 Diagnostic Runs

For the diagnostic runs listed in Table 2, the run names consist of a language code ("FI" for Finnish, "FR" for French, "PT" for Portuguese and "RU" for Russian) followed by one of the following labels:

- "lex": The run used SearchServer's lexical stemming with decompounding enabled, i.e. SET TERM_GENERATOR 'word!ftelp/inflect/decompound'. (Of the investigated languages, decompounding only makes a difference for Finnish.)
- "compound" (Finnish only): Same as "lex" except that /decompound was not specified.
- "single": Same as "lex" except that /single was additionally specified (so that just one stemming interpretation was used).

Table 2: Scores of Diagnostic Title-only runs

Run	AvgP	Robust@1	Robust@5	Robust@10
FI-lex	0.561	32/45 (71%)	36/45 (80%)	38/45 (84%)
FI-chain	0.553	30/45 (67%)	36/45 (80%)	39/45 (87%)
FI-single	0.550	32/45 (71%)	35/45 (78%)	37/45 (82%)
FI-compound	0.469	28/45 (62%)	30/45 (67%)	33/45 (73%)
FI-alg	0.424	26/45 (58%)	30/45 (67%)	34/45 (76%)
FI-none	0.328	19/45 (42%)	26/45 (58%)	27/45 (60%)
RU-lex	0.430	19/34 (56%)	27/34 (79%)	27/34 (79%)
RU-chain	0.405	18/34 (53%)	26/34 (76%)	26/34 (76%)
RU-single	0.396	17/34 (50%)	26/34 (76%)	27/34 (79%)
RU-alg	0.410	18/34 (53%)	26/34 (76%)	26/34 (76%)
RU-none	0.220	9/34 (26%)	20/34 (59%)	22/34 (65%)
PT-lex	0.405	24/46 (52%)	33/46 (72%)	35/46 (76%)
PT-chain	0.411	24/46 (52%)	34/46 (74%)	35/46 (76%)
PT-single	0.388	22/46 (48%)	31/46 (67%)	36/46 (78%)
PT-alg	0.387	25/46 (54%)	33/46 (72%)	34/46 (74%)
PT-none	0.327	18/46 (39%)	26/46 (57%)	31/46 (67%)
FR-lex	0.422	25/49 (51%)	39/49 (80%)	44/49 (90%)
FR-chain	0.418	26/49 (53%)	38/49 (78%)	42/49 (86%)
FR-single	0.423	26/49 (53%)	39/49 (80%)	44/49 (90%)
FR-alg	0.417	26/49 (53%)	38/49 (78%)	43/49 (88%)
FR-none	0.361	22/49 (45%)	39/49 (80%)	42/49 (86%)

- "alg": The run used a different index based on the coarser algorithmic Porter "Snowball" stemmer [5] for the language. Decompounding is not available with this stemmer and the /single option is redundant.
- "chain": The run used a different index based on applying the SearchServer stemmer (as "lex") and then the algorithmic stemmer.
- "none": The run disabled morphological searching, i.e. SET TERM_GENERATOR ''.

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

2.5 Evaluation Measures

The primary evaluation measure in this paper is "mean average precision" based on the first 1000 retrieved documents for each topic (denoted "AvgP" in Tables 2 and 9). "Average precision" for a topic is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant documents which are not retrieved). The score ranges from 0.0 (no relevants found) to 1.0 (all relevants found at the top of the list). For a set of topics, all topics are weighted equally by the mean. Average precision takes into account both precision and recall, and it is very good for detecting retrieval differences because even small differences in the ranks of relevant documents affect the score.

A more experimental measure is "robustness at 10 documents" (denoted "Robust@10") which is the percentage of topics for which at least one relevant document was returned in the first 10 rows (this was one of the measures investigated in the TREC Robust Retrieval track last year [11]). This measure hides a lot of retrieval differences (particularly in recall), but it may be an indicator of a user's impression of a method's robustness across topics. We also list the Robust@1 and Robust@5 variants.

Table 3: Impact of Lexical Stemming on Average Precision

Expt	AvgDiff	95% Conf	vs.	3 Extreme Diffs (Topic)
FI lex-none	0.233	(0.146, 0.326)	31-9-5	1.00 (224), 0.96 (210), −0.24 (208)
RU lex-none	0.209	(0.108, 0.325)	22-1-11	1.00 (250), 1.00 (203), −0.04 (228)
PT lex-none	0.078	(0.037, 0.125)	25-8-13	0.61 (213), 0.53 (229), −0.08 (248)
FR lex-none	0.061	(0.030, 0.096)	23-20-6	0.42 (229), 0.40 (235), −0.07 (216)

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. The difference is the first run minus the second run. For example, "FI lex-none" specifies the difference of subtracting the scores of the Finnish 'none' run from the Finnish 'lex' run (of Table 2).
- "AvgDiff" is the difference of the mean scores of the two runs being compared (the table heading says which evaluation measure is being compared).
- "95% Conf" is an approximate 95% confidence interval for the difference calculated using Efron's bootstrap percentile method[2] [2] (using 100,000 iterations). 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, 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 (45 for Finnish, 49 for French, 46 for Portuguese, 34 for Russian).
- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets (the topic numbers range from 201 to 250). 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 Morphological Experiments

This section looks at the differences between the runs of Table 2 in more detail.

3.1 Impact of Lexical Stemming

Table 3 isolates the impact of SearchServer's lexical stemming on the average precision measure (e.g. "FI lex-none" is the difference of the "FI-lex" and "FI-none" runs of Table 2). For each of the 4 languages, the increase in mean average precision was statistically significant (i.e. zero is not in the approximate 95% confidence interval). Note that for some queries, it is still better to only match the original query form (not inflections); SearchServer allows this option to be controlled for each query term at search-time.

Table 4 isolates the impact of SearchServer's lexical stemming on the Robust@10 measure. For each language, at least one relevant was found in the first 10 rows more often with inflections enabled than disabled; the increases were statistically significant for Finnish and Russian.

[2]See [9] for some comparisons of confidence intervals from the bootstrap percentile, Wilcoxon signed rank and standard error methods for both average precision and Precision@10.

Table 4: Impact of Lexical Stemming on Robustness at 10 Documents

Expt	AvgDiff	95% Conf	vs.	3 Extreme Diffs (Topic)
FI lex-none	0.244	(0.133, 0.378)	11-0-34	1.00 (224), 1.00 (215), 0.00 (250)
RU lex-none	0.147	(0.029, 0.265)	5-0-29	1.00 (221), 1.00 (244), 0.00 (226)
PT lex-none	0.087	(−0.001, 0.196)	5-1-40	1.00 (243), 1.00 (234), −1.00 (235)
FR lex-none	0.041	(−0.041, 0.123)	3-1-45	1.00 (241), 1.00 (239), −1.00 (222)

Table 5: Lexical vs. Algorithmic Stemming on Average Precision

Expt	AvgDiff	95% Conf	vs.	3 Extreme Diffs (Topic)
FI lex-alg	0.137	(0.064, 0.219)	26-12-7	0.98 (210), 0.86 (226), −0.15 (219)
RU lex-alg	0.019	(−0.003, 0.050)	7-8-19	0.40 (227), 0.20 (202), −0.07 (224)
PT lex-alg	0.018	(−0.014, 0.055)	16-14-16	0.53 (229), 0.32 (217), −0.27 (204)
FR lex-alg	0.005	(−0.003, 0.013)	18-14-17	−0.09 (203), 0.06 (209), 0.08 (231)

3.2 Comparison with Algorithmic Stemming

Table 5 contains the results of a diagnostic experiment comparing average precision when the only difference is the stemmer used: the experimental SearchServer lexical stemmer or Porter's algorithmic stemmer. Positive differences indicate that the SearchServer stemmer led to a higher score and negative differences indicate that the algorithmic stemmer led to a higher score. Using SearchServer's stemmer scored higher on average for each language and this increase was statistically significant for Finnish.

In this section, we look at the Portuguese and Russian topics with the largest differences. Finnish is examined in more detail in the subsequent decompounding section. French was investigated in last year's paper [10].

3.2.1 Portuguese Stemming

Topic PT-229: Table 5 shows that the largest difference between the stemming approaches for Portuguese was on topic 229 (Construção de Barragens (Dam Building)) in which average precision was 53 points higher with SearchServer's stemmer. The main reason was that, unlike the algorithmic stemmer, the SearchServer stemmer matched "Barragem", an inflection used in many relevant documents. SearchServer additionally matched "construções" which may also have been helpful.

Topic PT-217: The next largest difference for Portuguese was on topic 217 (Sida em África (AIDS in Africa)) for which Table 5 shows that average precision was 32 points higher with SearchServer's stemmer. The main reason was that, unlike SearchServer, the algorithmic stemmer matched "sido", a common word unrelated to AIDS, which decreased precision substantially. SearchServer additionally matched "africanos" which may also have been helpful.

Topic PT-204: The largest negative difference was on topic 204 (Vítimas de Avalanches (Victims of Avalanches)) for which using the algorithmic stemmer scored 27 points higher. Both stemmers matched "Avalanche" but the algorithmic stemmer additionally matched "avalancha" which was the only variant used in 3 of the relevant documents. We should investigate this case further.

3.2.2 Russian Stemming

Topic RU-227: Table 5 shows that the largest difference between the stemming approaches for Russian was on topic 227 (Алтайская амазонка (Altai Ice Maiden)) for which average precision

Table 6: Decompounding Experiments (Finnish) on Average Precision

Expt	AvgDiff	95% Conf	vs.	3 Extreme Diffs (Topic)
FI lex-cmpd	0.092	(0.034, 0.162)	17-9-19	0.98 (210), 0.72 (226), −0.18 (219)
FI cmpd-none	0.141	(0.075, 0.214)	27-10-8	1.00 (224), 0.81 (204), −0.22 (208)
FI cmpd-alg	0.045	(−0.001, 0.094)	19-14-12	0.57 (204), 0.50 (216), −0.40 (205)

was 40 points higher with SearchServer's stemmer. SearchServer internally produced 2 stems for "Алтайская" (Altai), itself and "Алтайскай". The words which had "Алтайская" as a stem (such as "Алтайской", "Алтайские", "Алтайскую" and "алтайских") were less common in the documents than the words which shared the "Алтайскай" stem (the same words plus more such as "Алтайского", "Алтайском" and "Алтайскому"), so SearchServer's experimental new scoring scheme for alternative stems gives the former group a higher weight from inverse document frequency than the latter group. In this case, it turned out just 1 relevant document was matched by either stemmer and it just used the original word "Алтайская". The algorithmic stemmer produced just one stem for these words, so its weighting did not have a preference for the query form and some documents with the second group of terms ended up ranking higher. The algorithmic stemmer additionally matched "Алтайске" which was not helpful in this case. This topic illustrates a benefit from SearchServer's experimental new handling of multiple stemming interpretations.

Topic RU-202: The next largest difference was on topic 202 (Арест Ника Леесон (Nick Leeson's Arrest)) for which the score was 20 points higher with SearchServer's stemmer. The 3 relevant documents used different spellings for "Leeson" ("Лисон", "Лизона", "Лизон" and "Лисона") which did not match the query form of "Леесон" with either stemmer. And inflections of "Арест" (Arrest) did not appear in the relevant documents. So the matches just came from variants of "Nick". Both stemmers matched the forms used in the relevant documents ("Ника" and "Ник"). But the algorithmic stemmer additionally matched other terms such as "Никому" and "никого" which lowered precision substantially in this case.

3.3 Impact of Decompounding (Finnish)

The first row of Table 6 ("FI lex-cmpd") isolates the impact of SearchServer's experimental new "/decompound" option for Finnish (decompounding is not new to SearchServer for Finnish, but an option to control its impact separately from inflectional stemming at search-time is). This option allows words to match if they share any stem of query compound words. Without the /decompound option, the (experimental new) default is to require all the stems of a compound word to be in the same or consecutive words to be considered a match. Table 6 shows that mean average precision was 9 points higher with /decompound set, and this difference was statistically significant.

The second row of Table 6 ("FI cmpd-none") shows that even without the /decompound option, use of SearchServer's stemming for Finnish scored 14 points higher than not using stemming. (Note that the first two rows of Table 6 add up to the 23 point gain from lexical stemming shown in Table 3.)

The third row of Table 6 ("FI cmpd-alg") compares SearchServer's stemming without the /decompound option to algorithmic stemming (which does not even decompound at index-time) and shows that using SearchServer's stemmer scored 4.5 points higher, though this difference did not quite pass the statistical significance test. (SearchServer's stemming with the /decompound option is compared to algorithmic stemming in Table 5 in which the difference is the sum of the differences of rows 1 and 3 of Table 6.)

We look at some Finnish topics in more detail to understand these results better.

Table 7: Impact of Adding Algorithmic to Lexical Stemming on Average Precision

Expt	AvgDiff	95% Conf	vs.	3 Extreme Diffs (Topic)
FI chain-lex	−0.008	(−0.037, 0.019)	13-16-16	−0.38 (203), −0.30 (215), 0.31 (205)
RU chain-lex	−0.025	(−0.062, 0.000)	5-8-21	−0.50 (226), −0.20 (202), 0.02 (232)
PT chain-lex	0.006	(−0.007, 0.023)	9-14-23	0.27 (204), 0.08 (205), −0.15 (232)
FR chain-lex	−0.004	(−0.010, 0.003)	8-18-23	0.09 (203), −0.06 (220), −0.06 (209)

3.3.1 Finnish Decompounding

Topic FI-210: Table 6 shows that the largest impact of Finnish decompounding was on topic 210 (Nobel rauhanpalkintoehdokkaat (Nobel Peace Prize Candidates)) for which using SearchServer's stemmer with the /decompound option scored 98 points higher than not using /decompound (and also 98 points higher than using the algorithmic stemmer according to Table 5). This topic had just 1 relevant document, and the only match for the non-decompounding approaches was the word "Nobel" which occurred in lots of documents, so the relevant document did not stand out among them. With SearchServer's decompounding, many more words in the relevant document matched such as "rauhan", "rauhanpalkituksi", "rauhanpalkinnon", "rauhanvälittäjänä", "ehdokasta" and "ehdokkaina" because these words shared at least one (but not all) the stems of the query compound "rauhanpalkintoehdokkaat", and the relevant document was ranked first.

Topic FI-226: Table 6 shows that the next largest impact of Finnish decompounding was on topic 226 (Sukupuolenvaihdosleikkaukset (Sex-change Operations)) for which using SearchServer's stemmer with the /decompound option scored 72 points higher than not using /decompound (and also 86 points higher than using the algorithmic stemmer according to Table 5). The algorithmic stemmer just found the one of the 13 relevant documents which contained the query word "Sukupuolenvaihdosleikkaukset". SearchServer without /decompound matched that document plus 3 other relevants, two which contained "sukupuolen vaihdosleikkaukseen" (an example of a consecutive-word match) and one which contained "Sukupuolenvaihdosleikkausta". SearchServer with /decompound matched all 13 relevant documents; the key additional matches appeared to be "Sukupuolen-vaihdos", "sukupuolenvaihtoleikkaukset", "sukupuolenvaihdot", "Sukupuolenvaihdoshan", "sukupuolenkorjausleikkausten" and "sukupuolenvahvistusleikkaus", though other matching words may also have been helpful such as "leikkaussali", "sukupuoli" and "vaihdos".

Topic FI-219: Table 6 shows that the largest negative impact of Finnish decompounding was on topic 219 (EU:n komissaariehdokkaat (EU Commissioner Candidates)) for which using SearchServer's stemmer with the /decompound option scored 18 points lower than not using /decompound (and also 15 points lower than using the algorithmic stemmer according to Table 5). Without the /decompound option, SearchServer found a lot of precise matches in relevant documents such as "komissaariehdokasta", "komissaariehdokkaalle", "komissaariehdokkaista", "komissaariehdokkaalta", "komissaariehdokkaiden" and "komissaariehdokkaan". Furthermore, in some relevant documents it found matches in larger compounds (which the algorithmic stemmer could not) such as "naiskomissaariehdokasta" and "tanskalaiseltakomissaariehdokkaalta". With /decompound set, SearchServer would also find all these matches, but precision was substantially hurt in this case by additionally matching terms in non-relevant documents such as "jäsenehdokkaiden", "jäsenehdokkaita", "ykkösehdokkaista", "tutkimuskomissaari" and "henkilöstökomissaari". This topic shows why a user may prefer to have /decompound not set; in cases where the user does not need the component words to occur together, the user can either manually separate the terms or set the /decompound option. But for automatic ad hoc searches for topics, it is better on average to use the /decompound option.

Table 8: Impact of Using All Lexical Stems on Average Precision

Expt	AvgDiff	95% Conf	vs.	3 Extreme Diffs (Topic)
FI lex-sing	0.010	(−0.002, 0.032)	7-10-28	0.44 (215), 0.05 (236), −0.02 (233)
RU lex-sing	0.033	(0.002, 0.082)	9-3-22	0.67 (203), 0.31 (210), −0.01 (233)
PT lex-sing	0.017	(−0.004, 0.049)	6-11-29	0.60 (213), 0.16 (236), −0.12 (248)
FR lex-sing	−0.001	(−0.004, 0.003)	2-9-38	0.06 (235), −0.03 (248), −0.03 (215)

3.4 Adding Algorithmic to Lexical Stemming

Table 7 shows the impact of applying the algorithmic stemmer to the result of SearchServer's stemmer (this is possible because SearchServer's stemmer returns real words; the other order would not work because the algorithmic stemmer often truncates to a non-word). This approach would still produce all the matches of SearchServer's stemming and may sometimes produce additional matches from algorithmic stemming. However, there was a decrease in mean average precision for Russian which was borderline significant. The other differences were not statistically significant. While algorithmic stemming may occasionally add a helpful match, it can also add poor matches that hurt precision. In a future experiment, perhaps it would be better to treat algorithmic stems as alternative stemming interpretations (instead of replacing the lexical stem) so that lexical inflections are likely to get higher weight when the algorithmic stem is too common.

3.5 Impact of Using All Lexical Stems

Table 8 shows the impact of using all stemming interpretations from SearchServer's lexical stemming instead of arbitrarily just using the first one. The increase in mean average precision was statistically significant for Russian. On the individual topics, there were some large increases, but (reassuringly) no correspondingly large decreases. In past years, mean average precision was typically lower when including all the stems because of over-weighting issues, so this result suggests that the enhancement for handling multiple stemming interpretations has succeeded at addressing this issue.

Topic RU-203: The largest difference for Russian was on topic 203 (Партизанская война в Восточном Тиморе (East Timor Guerrillas)) for which the score was 67 points higher when using all stemming interpretations. The query word "Восточном" (Eastern) had 2 stems, "Восточнома" and "восточный". The inflections in the relevant document (namely "Восточного", "Восточный", "восточных" and "восточной") only shared the latter stem.

4 Submitted Runs

In the identifiers of the runs submitted for assessment in May 2004 (e.g. "humFI04tde"), the first 3 letters "hum" indicate a Hummingbird submission, the next 2 letters are the language code, and the number "04" indicates CLEF 2004. "t", "d" and "n" indicate that the Title, Description and Narrative field of the topic were used (respectively). "e" indicates that query expansion from blind feedback on the first 2 rows was used (see last year's paper [10] for more details). The submitted runs all used inflections from SearchServer's lexical stemming (including decompounding where applicable). The scores of the submitted runs are listed in Table 9.

The submitted Title-only runs (e.g. "humFI04t" of Table 9) correspond to the "lex" diagnostic runs (e.g. "FI-lex" of Table 2) except that the submitted runs used an older experimental version of SearchServer (including an older version of the lexical stemming component) so the scores are not exactly the same.

Table 9: Scores of Submitted Runs

Run	AvgP	Robust@1	Robust@5	Robust@10
humFI04t	0.556	34/45 (76%)	35/45 (78%)	37/45 (82%)
humFI04td	0.593	31/45 (69%)	37/45 (82%)	38/45 (84%)
humFI04tde	0.637	32/45 (71%)	40/45 (89%)	42/45 (93%)
humRU04t	0.430	19/34 (56%)	27/34 (79%)	27/34 (79%)
humRU04td	0.409	17/34 (50%)	26/34 (76%)	27/34 (79%)
humRU04tde	0.443	17/34 (50%)	26/34 (76%)	27/34 (79%)
humPT04t	0.405	24/46 (52%)	33/46 (72%)	35/46 (76%)
humPT04td	0.453	23/46 (50%)	32/46 (70%)	34/46 (74%)
humPT04tde	0.475	23/46 (50%)	32/46 (70%)	35/46 (76%)
humFR04t	0.421	25/49 (51%)	39/49 (80%)	44/49 (90%)
humFR04td	0.458	26/49 (53%)	43/49 (88%)	44/49 (90%)
humFR04tde	0.493	26/49 (53%)	43/49 (88%)	43/49 (88%)

References

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

[2] Bradley Efron and Robert J. Tibshirani. An Introduction to the Bootstrap. 1993. Chapman & Hall/CRC.

[3] Andrew Hodgson. Converting the Fulcrum Search Engine to Unicode. In Sixteenth International Unicode Conference, Amsterdam, The Netherlands, March 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, M. Gatford. (City University.) Okapi at TREC-3. In D. K. Harman, editor, Overview of the Third Text REtrieval Conference (TREC-3). NIST Special Publication 500-226. http://trec.nist.gov/pubs/trec3/t3_proceedings.html

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

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

[9] Stephen Tomlinson. Experiments in 8 European Languages with Hummingbird SearchServer™ at CLEF 2002. In Carol Peters, editor, Working Notes for the CLEF 2002 Workshop. http://clef.iei.pi.cnr.it:2002/workshop2002/WN/26.pdf

[10] Stephen Tomlinson. Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer™ at CLEF 2003. In Carol Peters, editor, Working Notes for the CLEF 2003 Workshop. http://clef.iei.pi.cnr.it/2003/WN_web/19.pdf

[11] Ellen M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2003). http://trec.nist.gov/pubs/trec12/t12_proceedings.html