Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer™ at CLEF 2003

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

July 20, 2003

Abstract

Hummingbird participated in the monolingual information retrieval tasks of the Cross-Language Evaluation Forum (CLEF) 2003: for natural language queries in 9 European languages (German, French, Italian, Spanish, Dutch, Finnish, Swedish, Russian and English) find all the relevant documents (with high precision) in the CLEF 2003 document sets. For each language, SearchServer scored higher than the median average precision on more topics than it scored lower. In a comparison of experimental SearchServer lexical stemmers with Porter's algorithmic stemmers, the biggest differences were for the languages in which compound words are frequent (German, Dutch, Finnish and Swedish). SearchServer scored significantly higher in average precision for German and Finnish, apparently from its ability to split compound words and find terms when they are parts of compounds in these languages. Most of the differences for the other languages appeared to be from SearchServer's lexical stemmers performing inflectional stemming while the algorithmic stemmers often additionally performed derivational stemming; these differences did not pass a significance test.

1 Introduction

Hummingbird SearchServer[1] is an indexing, search and retrieval engine for embedding in Windows and UNIX information applications. SearchServer, originally a product of Fulcrum Technologies, was acquired by Hummingbird in 1999. Founded in 1983 in Ottawa, Canada, Fulcrum produced the first commercial application program interface (API) for writing information retrieval applications, Fulcrum® Ful/Text™. The SearchServer kernel is embedded in many Hummingbird products, including SearchServer, an application toolkit used for knowledge-intensive applications that require fast access to unstructured information.

SearchServer supports a variation of the Structured Query Language (SQL), SearchSQL™, which has extensions for text retrieval. SearchServer conforms to subsets of the Open Database Connectivity (ODBC) interface for C programming language applications and the Java Database Connectivity (JDBC) interface for Java applications. Almost 200 document formats are supported, such as Word, WordPerfect, Excel, PowerPoint, PDF and HTML.

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

[1]Fulcrum® is a registered trademark, and SearchServer™, SearchSQL™, Intuitive Searching™ and Ful/Text™ are trademarks of Hummingbird Ltd. All other copyrights, trademarks and tradenames are the property of their respective owners.

Table 1: Sizes of CLEF 2003 Document Sets

Language	Text Size (uncompressed)	Number of Documents
Spanish	1,158,177,739 bytes (1105 MB)	454,045
German	704,523,506 bytes (672 MB)	294,809
Dutch	558,560,087 bytes (533 MB)	190,604
English	601,737,745 bytes (574 MB)	169,477
Italian	378,831,019 bytes (361 MB)	157,558
Swedish	374,371,465 bytes (357 MB)	142,819
French	344,961,357 bytes (329 MB)	129,806
Finnish	143,902,109 bytes (137 MB)	55,344
Russian	68,802,653 bytes (66 MB)	16,716

TREC [7]) have provided opportunities to objectively evaluate SearchServer's support for natural language queries in more than a dozen languages.

This (draft) paper looks at experimental work with SearchServer for the task of finding relevant documents for natural language queries in 9 European languages using the CLEF 2003 test collections. For the experiments described in this paper, an experimental post-5.x version of SearchServer was used.

2 Methodology

2.1 Data

The CLEF 2003 document sets consisted of tagged (SGML-formatted) news articles (mostly from 1994 and 1995) in 9 different languages: German, French, Italian, Spanish, Dutch, Swedish, Finnish, Russian and English. Compared to last year, Russian was new, and there were more documents in Spanish, German, Italian, French and English. The English documents included some British English for the first time. Table 1 gives the sizes.

The CLEF organizers created 60 natural language "topics" (numbered 141-200) and translated them into many languages. Each topic contained a "Title" (subject of the topic), "Description" (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.

For more information on the CLEF test collections, see the CLEF web site [1].

2.2 Indexing

A separate SearchServer table was created for the documents of each language. For details of the SearchServer syntax, see last year's paper [8].

Unlike last year, we used SearchServer's default of not indexing accents for all languages, except for Russian, for which we indexed the combining breve (Unicode 0x0306) so that the Cyrillic Short I (0x0419) was not normalized to the Cyrillic I (0x0418).

We treated the apostrophe as a word separator for all languages except English.

Typically, a couple hundred stop words were excluded from indexing for each language (e.g. "the", "by" and "of" in English). The Porter web site [5] contains stop word lists for most European languages. We used its list for Russian, but our lists for other languages may contain differences.

SearchServer internally uses Unicode. A different option to SearchServer's translation text reader was specified for Russian (UTF8_UCS2) than for the other languages (Win_1252_UCS2) because the Russian documents were encoded in the UTF-8 character set and the documents for

the other languages were encoded in the Latin-1 character set. (A custom text reader, cTREC, was also updated to maintain support for the CLEF guidelines of only indexing specifically tagged fields; the new British and Russian collections necessitated the update.)

By default, the SearchServer index supports both exact matching (after some Unicode-based normalizations, such as converting to upper-case and decomposed form) and matching of inflections.

2.3 Lexical Stemming

For many languages (including all 9 European languages of CLEF 2003), 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 experimental development version of SearchServer used for the experiments in this paper was based on Inxight LinguistX Platform 3.5. Unlike the previous two years, the lexical stemming was conducted in an "expanded" mode which tolerates missing accents (e.g. unlike last year, "bebes" stems to "bébé" in French) and handles more plural cases (e.g. unlike last year, "PCs" stems to "PC" in English).

For all languages, we used inflectional stemming which generally retains the part of speech (e.g. a plural of a noun is typically stemmed to the singular form). We did not use 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 German, Dutch and Finnish (but not for Swedish in this version, and not for the other languages as it is not generally applicable). 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".

2.4 Intuitive Searching

For all runs, we used SearchServer's Intuitive Searching, i.e. the IS_ABOUT predicate of SearchSQL, which accepts unstructured natural language text. For example, for the German version of topic 41 (from a previous year), the Title was "Pestizide in Babykost" (Pesticides in Baby Food), and the Description was "Berichte über Pestizide in Babynahrung sind gesucht" (Find reports on pesticides in baby food). A corresponding SearchSQL query would be:

```
SELECT RELEVANCE('V2:3') AS REL, DOCNO
FROM CLEF03DE
WHERE FT_TEXT IS_ABOUT 'Pestizide in Babykost Berichte über
Pestizide in Babynahrung sind gesucht'
ORDER BY REL DESC;
```

For the Russian queries, the statement "SET CHARACTER_SET 'UTF8C' " was previously executed because the queries were in UTF-8 instead of Latin-1.

2.5 Statistical Relevance Ranking

SearchServer's relevance value calculation is the same as described last year [8]. 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. SearchServer's relevance values are always an integer in the range 0 to 1000.

SearchServer's RELEVANCE_METHOD setting can be used to optionally square the importance of the inverse document frequency (by choosing a RELEVANCE_METHOD of 'V2:4' instead of 'V2:3'). The importance of document length to the ranking is controlled by SearchServer's RELEVANCE_DLEN_IMP setting (scale of 0 to 1000). For all runs in this paper, RELEVANCE_METHOD was set to 'V2:3' and RELEVANCE_DLEN_IMP was set to 750.

2.6 Query Stop Words

We automatically removed words such as "find", "relevant" and "document" from the topics before presenting them to SearchServer, i.e. words which are not stop words in general but were commonly used in the CLEF topics as general instructions. For the submitted runs, the lists were developed by examining the CLEF 2000, 2001 and 2002 topics (not this year's topics). An evaluation in last year's paper [8] found this step to be of only minor impact.

2.7 Query Expansion

For one of the submitted runs for each language (the runs with identifiers ending with 'e', e.g. humDE03tde), the first 3 rows from the other submitted run for the language (e.g. humDE03td) were used to find additional query terms. Only terms appearing in at most 5% of the documents (based on the most common inflection of the term) were included. Mathematically, the approach is similar to Rocchio feedback with weights of one-half for the original query and one-sixth for each of the 3 expansion rows. See section 5.2 of [9] for more details. This is the first time we have used a blind feedback technique for CLEF submissions. We did not use it for any of the diagnostic experiments.

2.8 Evaluation Measures

The evaluation measures are likely explained in an appendix of this volume. Briefly: "Precision" is the percentage of retrieved documents which are relevant. "Precision@n" is the precision after n documents have been retrieved. "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). "Recall" is the percentage of relevant documents which have been retrieved. "Interpolated precision" at a particular recall level for a topic is the maximum precision achieved for the topic at that or any higher recall level. For a set of topics, the measure is the mean of the measure for each topic (i.e. all topics are weighted equally).

The Monolingual Information Retrieval tasks were to run 60 queries against document collections in the same language and submit a list of the top-1000 ranked documents to CLEF for judging (in May 2003). CLEF produced a "qrels" file for each of the 9 tasks: a list of documents judged to be relevant or not relevant for each topic. (For Swedish, this draft paper uses the preliminary set of qrels.)

For some topics and languages, no documents were judged relevant. The precision scores are just averaged over the number of topics for which at least one document was judged relevant.

For tables focusing on the impact of one particular difference in approach (such as a stemming method as in Table 2), the columns are as follows:

- "Experiment" is the language and topic fields used (for example, "-td" indicates the Title and Description fields were used).
- "AvgDiff" is the average (mean) difference in the precision score.
- "95% Confidence" is an approximate 95% confidence interval for the average 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

[2]See last year's paper [8] for some comparisons of confidence intervals from the bootstrap percentile, Wilcoxon signed rank and standard error methods for both average precision and Precision@10.

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 precision was higher, lower and tied (respectively) with the feature enabled. These numbers should always add to the number of topics for the language (as per Table 3).
- "2 Largest Diffs (Topic)" lists the two largest differences in the precision score (based on the absolute value), with each followed by the corresponding topic number in brackets (the topic numbers range from 141 to 200).

For tables providing multiple precision scores (such as Table 3), listed for each run are its mean average precision (AvgP), the mean precision after 5, 10 and 20 documents retrieved (P@5, P@10 and P@20 respectively), the mean interpolated precision at 0% and 30% recall (Rec0 and Rec30 respectively), and the mean precision after R documents retrieved (P@R) where R is the number of relevant documents for the topic. The number of topics with at least one relevant document is also included in this table, though it is a property of the test collection, not of the run.

2.9 Submitted Runs

In the identifiers for the submitted runs (e.g. humDE03tde), the first 3 letters "hum" indicate a Hummingbird submission, the next 2 letters are the language code, and the number "03" indicates CLEF 2003. "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 was used. The submitted runs all used inflections from SearchServer's lexical stemming.

The following language codes were used: "DE" for German, "EN" for English, "ES" for Spanish, "FI" for Finnish, "FR" for French, "IT" for Italian, "NL" for Dutch, "RU" for Russian, and "SV" for Swedish.

For each language, we submitted a "td" and "tde" run (namely "humDE03td", "humDE03tde", "humFR03td", "humFR03tde", "humIT03td", "humIT03tde", "humES03td", "humES03tde", "humNL03td", "humNL03tde", "humFI03td", "humFI03tde", "humSV03td", "humSV03tde", "humRU03td" and "humRU03tde"). Note that monolingual English submissions were not allowed. For Russian, additional runs were requested for the judging pools, so we also submitted Title-only runs ("humRU03t" and "humRU03te") and full topic runs ("humRU03tdn" and "humRU03tdne"). For 3 other Russian submissions ("humRU03tm", "humRU03tdm", "humRU03tdnm"), the "m" was meant to indicate that morphology (stemming) was disabled, but by accident for these runs the CHARACTER_SET was set to Latin-1 instead of UTF-8, which led to precision scores of almost zero.

The scores of the submitted runs are likely listed in an appendix of this volume.

3 Comparison of Lexical and Algorithmic Stemming

The experimental version of SearchServer used for these experiments allows plugging-in of custom stemming modules. As a test for this feature, we have experimented with plugging-in Porter's algorithmic "Snowball" stemmers [5]. For English, the Porter2 version was used.

Table 2 contains the results of a diagnostic experiment comparing average precision for the short (Title-only) queries 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. SearchServer's stemmer scored significantly higher for Finnish and German and significantly lower for Swedish. The differences for the other languages didn't pass the significance test.

To try to better understand the differences between these approaches to stemming, we look at least at the topics for each language with the two biggest differences in the average precision

Table 2: Lexical vs. Algorithmic Stemming for Average Precision, Title-only queries

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
FI-stem-t	0.131	(0.032, 0.231)	28-14-3	−0.998 (185), 0.929 (196)
DE-stem-t	0.104	(0.054, 0.159)	39-13-4	0.833 (174), 0.596 (158)
NL-stem-t	0.035	(−0.009, 0.082)	28-20-8	0.635 (174), 0.494 (165)
RU-stem-t	0.018	(−0.046, 0.098)	10-8-10	0.800 (187), 0.338 (177)
ES-stem-t	0.005	(−0.008, 0.017)	29-14-14	−0.183 (186), 0.170 (151)
FR-stem-t	−0.004	(−0.027, 0.017)	18-14-20	−0.359 (145), 0.254 (177)
EN-stem-t	−0.005	(−0.025, 0.019)	13-23-18	0.469 (180), −0.225 (179)
IT-stem-t	−0.028	(−0.078, 0.006)	22-18-11	−1.000 (161), −0.287 (157)
SV-stem-t	−0.030	(−0.060,−0.005)	14-24-15	−0.500 (188), −0.333 (144)

score (usually we look at more than two). We just look at the shorter Title-only topics for ease of analysis (fewer words in the query makes it easier to see what caused the difference) and because shorter queries are preferred by users anyway.

3.1 English Stemming

English topics 180 (Bankruptcy of Barings), 179 (Resignation of NATO Secretary General), 175 (Everglades Environmental Damage) and 168 (Assassination of Rabin) show that the algorithmic stemmer often performs derivational stemming (whereas the SearchServer stemmer is known to just do inflectional stemming as described earlier). In the case of topic 180, derivational stemming lowered the average precision score because it was harmful for this topic to match "Barings" with "bare", "bares" and "barely". But for topics 179, deriving "resign" and "resigned" from "resignation" was apparently helpful. Likewise, for topic 175, deriving "environment" from "environmental" was apparently helpful, and in topic 168 deriving "assassin" from "assassination" was apparently helpful. SearchServer's stemmer internally has the option of derivational stemming for English (and handles all of these cases similarly), but there is not currently an option to enable it. It might make for an interesting future experiment to try it.

English topic 200 (Flooding in Holland and Germany) illustrated that another difference for English is the handling of apostrophe-S. Perhaps surprisingly, the algorithmic stemmer never removes apostrophe-S. The SearchServer stemmer does remove it in some cases, e.g. it appears SearchServer scored higher on topic 200 because it matched "Holland's" with "Holland" and "Germany's" with "Germany". In topic 169, SearchServer matched "NATO's" with "NATO" and "general's" with "general". But in topic 168, "Rabin's" was not matched with "Rabin", so SearchServer is not using a simple rule (a more familiar case is that SearchServer does not match "Parkinson's" to "Parkinson"). For the other languages, we treated the apostrophe as a word separator, so handling of apostrophes won't be an issue.

3.2 French Stemming

French topics 145 (Le Japon et ses importations de riz (Japanese Rice Imports)) and 177 (La consommation de lait en Europe (Milk Consumption in Europe)) illustrate that the French algorithmic stemmer also does some derivational stemming. In topic 145, the algorithmic stemmer matched the noun "imports" with verb forms such as "importé" and "importer", which apparently was helpful to the average precision score (though additionally deriving the unrelated terms "importance" and "important" might be disconcerting to a user). It also derived "Japonais" from "Japon". In topic 177, deriving "consommateurs" (consumers) and "consommateur" (consumer) from "consommation" (consumption) apparently hurt average precision.

French topic 162 (l'Union Européenne et les douanes turques (EU and Turkish Customs)) shows that sometimes SearchServer handles irregular inflections that the algorithmic stemmer does not.

SearchServer matched “turques” with “turc” and “turcs”, unlike the algorithmic stemmer. Both matched “turques” with “turque”. The algorithmic stemmer additionally derived “turquie” which appears to be why it scored higher on this topic. Overall, for the French topics, Table 2 shows that neither stemmer scored significantly higher than the other (the confidence interval contains zero).

3.3 Italian Stemming

In Italian topic 161 (Diete per Celiaci (Diets for Celiacs)), the algorithmic stemmer found the one relevant document by matching “celiaci” with “celiaca”. SearchServer stemmed “celiaci” to “celiare” and “celiaca” to itself and so did not make this match. We should investigate this case further.

In Italian topic 157 (Campionesse di Wimbledon (Wimbledon Lady Winners)), both stemmers matched “campionesse” with “campionessa”, but SearchServer additionally matched “campioni” and “campione”, which hurt average precision in this case.

In Italian topic 187 (Trasporto Nucleare in Germania (Nuclear Transport in Germany)), SearchServer scored higher, apparently from matching “nucleare” with “nucleari”, unlike the algorithmic stemmer.

3.4 Spanish Stemming

In Spanish topic 186 (Coalición del gobierno holandés (Dutch Coalition Government)), SearchServer matched “holandés” with “holandeses” and “holandesa”, unlike the algorithmic stemmer, and SearchServer scored a good 0.57 average precision, but the algorithmic stemmer derived “holandés” to “holanda”, which apparently helped it score higher (0.75).

In Spanish topic 151 (Las maravillas del Mundo Antiguo (Wonders of Ancient World)), the algorithmic stemmer derived more terms from “maravillas” (wonders) such as “maravilloso” (wonderful) which hurt precision. Both stemmers matched “Antiguo” with “antiguos” and “antigua” (among others), and SearchServer additionally matched “antiquísima” which may have been helpful.

3.5 German Stemming

For German topic 174 (Bayerischer Kruzifixstreit (Bavarian Crucifix Quarrel)), SearchServer split the compound word “Kruzifixstreit” and found many relevant documents by matching terms such as “Kruzifix”, “Kruzifixen” and “Kruzifixe” (and also “Streit”, though it seemed less important in this case). The algorithmic stemmer does not support compound-splitting, and “Kruzifixstreit” did not itself appear in the document set (nor did any compound variant of it), so it scored dramatically lower for this topic as can be seen in Table 2.

For German topic 158 (Fußball-Rowdys in Dublin (Soccer Riots in Dublin)), even though there was no compound word in the query, the relevant documents used compound words such as “Fussballrowdies” and “Fussballfans” which SearchServer successfully matched but the algorithmic stemmer did not.

German topic 190 (Kinderarbeit in Asien (Child Labor in Asia)) shows that compound-splitting is not always helpful. In this topic it hurt precision a lot to split “Kinderarbeit”, presumably because the term was typically used in that form in the relevant documents, and a lot of other documents used the German words for children and work in other contexts. (This happens a lot in information retrieval; a technique that works well on average can still have a substantial percentage of cases for which it is harmful. While there may be room for automatic improvement, it’s a good idea for applications to let the user override the defaults when desired.)

Overall for German, Table 2 shows that the SearchServer stemmer scored significantly higher on average, presumably because of compound-splitting. It appears it would be hard to isolate the impact of other differences because even when none of the query terms are compound words, the terms in the documents may be parts of compounds.

3.6 Dutch Stemming

Dutch topic 174 (Beierse Kruisbeeldstrijd (Bavarian Crucifix Quarrel)) is the Dutch version of the crucifix query examined earlier for German. SearchServer scores highly for similar reasons, i.e. SearchServer splits the compound and matches "kruisbeeld" and "strijd" among other forms. "Kruisbeeldstrijd" did not itself appear in the document set and the algorithmic stemmer scored dramatically lower.

Dutch topic 165 (Golden Globes 1994 (Golden Globes 1994)) is a case for Dutch in which a large difference in average precision did not result from compound handling differences. SearchServer apparently scored higher from matching "Globes" with "Globe" and perhaps also from matching "golden" with "gelden". If compound words aren't as frequent in Dutch as German, that may be why the overall differences between the stemmers did not quite pass the significance test.

3.7 Finnish Stemming

For Finnish topic 185 (Hollantilaisten valokuvat Srebrenicasta (Dutch Photos of Srebrenica)), SearchServer did not match any of "Srebrenicassa", "Srebrenica" and "Srebrenican", variants of "Srebrenicasta" in the relevant document matched by the algorithmic stemmer. Srebrenica is a proper noun. Porter mentions in [5] that "in a language in which proper names are inflected (Latin, Finnish, Russian ...), a dictionary-based stemmer will need to remove i-suffixes independently of dictionary look-up, because the proper names will not of course be in the dictionary." We should investigate if we are handling proper nouns adequately for languages such as Finnish and Russian.

Finnish topic 196 (Japanilaisten pankkien fuusio (Merger of Japanese Banks)) also illustrates how inflective a language Finnish is. SearchServer matched several terms in the two relevant documents that the algorithmic stemmer did not such as "Japanilaisen", "Japaniin", "Japanilaiset", "japanilaisia", "japanilaispankin" (a compound) and "pankin", apparently helping it to score much higher.

Finnish topic 147 (Öljyonnettomuudet ja linnut (Oil Accidents and Birds)) is a case showing the importance of compounding to Finnish. SearchServer matched terms such as "Onnettomuuksien", "linturyhmä", "öljyonnettomuuksien", "lintuvahinkojen", "Öljykatastrofi", "öljy" and "lintuja" (just to name a few) which appeared to be missed by the algorithmic stemmer (though not all of these were from compound-splitting) and SearchServer scored substantially higher.

3.8 Swedish Stemming

Swedish topic 188 (Tysk stavningsreform (German Spelling Reform)) shows that when a lexicon-based stemmer does not support compound-splitting for a language with frequent compounds (which is currently the case for SearchServer regarding Swedish), a secondary penalty is that inflections of compounds can be missed. In this topic, SearchServer did not match "stavningsreform" to "stavningsreformen", even though it matches "reform" to "reformen", presumably because the lexicon does not contain most compound words. The algorithmic stemmer did match "stavningsreformen" which apparently is why it scored higher on this topic.

For Swedish topic 144 (Uppror i Sierra Leone och diamanter (Sierra Leone Rebellion and Diamonds)), it appears the difference in the score was from SearchServer matching "uppror" with "upproret" while the algorithmic stemmer did not. SearchServer's behaviour looks reasonable but it appears it was not helpful in this case just by chance (the top retrieved documents had similar relevance scores and the small shift caused by this difference happened to move down a relevant document).

Swedish topic 187 (Kärnavfallstransporter i Tyskland (Nuclear Transport in Germany)) is another case like topic 188. SearchServer did not match "Kärnavfallstransporter", a Swedish compound word, with "kärnavfallstransport" nor "kärnavfallstransporten" (even though SearchServer does match "transporter", "transport" and "transporten" with each other). The algorithmic stemmer handled all of these cases and scored higher on this topic.

Table 3: Precision with Lexical, Algorithmic and No Stemming, Title-only queries

Run	AvgP	P@5	P@10	P@20	Rec0	Rec30	P@R	Topics
FI-lex-t	0.553	47.6%	35.3%	26.0%	0.762	0.682	52.5%	45
FI-alg-t	0.422	37.8%	27.8%	21.0%	0.682	0.539	40.9%	45
FI-none-t	0.301	30.2%	23.8%	17.8%	0.555	0.398	29.1%	45
DE-lex-t	0.424	59.6%	51.1%	40.8%	0.780	0.557	42.5%	56
DE-alg-t	0.319	47.5%	40.2%	31.5%	0.666	0.402	32.9%	56
DE-none-t	0.267	44.6%	35.7%	27.8%	0.635	0.333	28.6%	56
RU-lex-t	0.315	25.7%	17.5%	11.1%	0.572	0.449	29.4%	28
RU-alg-t	0.297	28.6%	20.7%	13.2%	0.510	0.420	26.0%	28
RU-none-t	0.254	25.0%	17.5%	10.4%	0.493	0.389	23.1%	28
SV-lex-t	0.338	35.5%	26.0%	19.2%	0.665	0.439	32.6%	53
SV-alg-t	0.368	35.5%	27.4%	20.2%	0.706	0.487	36.5%	53
SV-none-t	0.286	31.3%	23.2%	17.3%	0.593	0.352	28.2%	53
NL-lex-t	0.422	45.4%	38.0%	32.1%	0.671	0.514	40.3%	56
NL-alg-t	0.388	44.6%	34.8%	29.0%	0.652	0.505	37.6%	56
NL-none-t	0.372	42.9%	33.9%	28.1%	0.649	0.487	37.1%	56
FR-lex-t	0.447	40.4%	31.5%	24.5%	0.689	0.549	41.5%	52
FR-alg-t	0.451	40.4%	31.7%	25.0%	0.672	0.559	41.1%	52
FR-none-t	0.413	38.1%	29.2%	23.3%	0.671	0.518	38.7%	52
ES-lex-t	0.405	51.9%	44.0%	36.1%	0.803	0.535	40.1%	57
ES-alg-t	0.400	50.9%	43.2%	35.9%	0.783	0.521	39.6%	57
ES-none-t	0.374	46.7%	42.6%	34.4%	0.762	0.494	37.4%	57
IT-lex-t	0.394	40.4%	30.2%	21.9%	0.683	0.487	36.2%	51
IT-alg-t	0.422	41.2%	30.8%	22.4%	0.727	0.526	40.0%	51
IT-none-t	0.367	35.7%	25.9%	19.7%	0.649	0.445	34.1%	51
EN-lex-t	0.448	38.5%	34.4%	27.8%	0.676	0.550	43.4%	54
EN-alg-t	0.453	38.5%	34.3%	27.2%	0.678	0.547	43.2%	54
EN-none-t	0.435	40.0%	32.4%	27.1%	0.676	0.542	42.8%	54

Table 4: Impact of Lexical Stemming on Average Precision, Title-only queries

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
FI-lex-t	0.252	(0.149, 0.360)	32-11-2	1.000 (147), 0.999 (187)
DE-lex-t	0.157	(0.103, 0.213)	43-10-3	0.843 (174), 0.627 (192)
RU-lex-t	0.062	(0.002, 0.146)	15-7-6	0.978 (187), 0.223 (143)
SV-lex-t	0.051	(0.023, 0.085)	23-15-15	0.507 (195), 0.479 (192)
NL-lex-t	0.050	(0.001, 0.102)	30-19-7	0.709 (174), 0.487 (188)
FR-lex-t	0.034	(−0.023, 0.091)	25-18-9	0.923 (175), −0.875 (141)
ES-lex-t	0.031	(0.011, 0.052)	33-19-5	0.240 (164), 0.228 (181)
IT-lex-t	0.027	(0.006, 0.050)	24-18-9	0.317 (171), 0.202 (200)
EN-lex-t	0.013	(−0.007, 0.038)	23-21-10	0.417 (144), 0.262 (158)

Swedish topic 179 (NATO:s generalsekreterares avsked (Resignation of NATO Secretary General)) is a case in which the opposite happened. SearchServer matched "generalsekreterares" with "generalsekreterare" while the algorithmic stemmer did not. Perhaps this word is handled because even though it looks like a compound, it probably is better not to split it because it has a different meaning as one word than it does if split in two. SearchServer scored higher on this topic.

Overall, Swedish is the one language (of the nine investigated) in which SearchServer's stemmer scored significantly lower than the algorithmic stemmer overall. Even though neither stemmer supports compound-splitting for Swedish, it appears for the lexicon-based stemmer this has a secondary penalty of causing inflections of some compounds to be missed. Adding compound-splitting support for Swedish would both overcome this issue and also allow terms to be found when they are parts of compounds.

3.9 Russian Stemming

For Russian, it appears the algorithmic stemmer tends to match more terms than SearchServer's stemmer, which might be a derivational vs. inflectional difference again, but we haven't investigated in detail yet. For Russian topics 187 and 177, SearchServer's stemmer scored higher, apparently from matching fewer forms of the Russian words for "nuclear" and "milk" respectively, which helped precision. The algorithmic stemmer scored higher on topic 148, apparently from matching more variations of the Russian word for "ozone". Overall, the differences did not pass a significance test.

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

[8] 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

[9] Stephen Tomlinson. Hummingbird SearchServer™ at TREC 2001. In E. M. Voorhees and D. K. Harman, editors, Proceedings of the Tenth Text REtrieval Conference (TREC 2001). NIST Special Publication 500-250. http://trec.nist.gov/pubs/trec10/t10_proceedings.html