Preprint of the publication: Kettunen, K. & Airio, E. & Järvelin, K. (2007). Restricted inflectional form generation in management of morphological keyword variation. Information Retrieval 10(4-5):415-444. (DOI 10.1007/s10791-007-9030-z).
.

Restricted Inflectional Form Generation in Management of Morphological Keyword Variation

Kimmo Kettunen, Eija Airio and Kalervo Järvelin

Dept. of Information Studies

University of Tampere, Finland

Kimmo.kettunen@uta.fi, Eija.airio@uta.fi, kalervo.jarvelin@uta.fi

Abstract

Word form normalization through lemmatization or stemming is a standard procedure in information retrieval because morphological variation needs to be accounted for and several languages are morphologically non-trivial. Lemmatization is effective but often requires expensive resources. Stemming is also effective in most contexts, generally almost as good as lemmatization and typically much less expensive; besides it also has a query expansion effect. However, in both approaches the idea is to turn many inflectional word forms to a single lemma or stem both in the database index and in queries. This means extra effort in creating database indexes.

In this paper we take an opposite approach: we leave the database index un-normalized and enrich the queries to cover for surface form variation of keywords. A potential penalty of the approach would be long queries and slow processing. However, we show that it only matters to cover a negligible number of possible surface forms even in morphologically complex languages to arrive at a performance that is almost as good as that delivered by stemming or lemmatization. Moreover, we show that, at least for typical test collections, it only matters to cover nouns and adjectives in queries. Furthermore, we show that our findings are particularly good for short queries that resemble normal searches of web users.

Our approach is called FCG (for Frequent Case (form) Generation). It can be relatively easily implemented for Latin/Greek/Cyrillic alphabet languages by examining their (typically very skewed) nominal form statistics in a small text sample and by creating surface form generators for the 3–9 most frequent forms. We demonstrate the potential of our FCG approach for several languages of varying morphological complexity: Swedish,

German, Russian, and Finnish in well-known test collections. Applications include in particular Web IR in languages poor in morphological resources.

1. Introduction

Various methods for handling the morphological variation of keywords in information retrieval (IR) have been used already for decades. Some of them are more complex than others, while some are amazingly simple but produce still quite good results in IR. So far it has been shown among other things that even a quite simple rule-based non-lexical stemmer can improve precision and recall of textual searches for languages that are morphologically more complex than English or some times even very complex – as, e.g., Finnish and Slovene (cf. Popovič and Willett 1992; Hollink et al. 2004; Airio 2006). Use of stemming has been a de facto standard in information retrieval, but in language technology use of full coverage lemmatization has been thought a necessity for languages that are morphologically complex, even in monolingual single term IR (Koskenniemi 1996). This belief has been shared also by some IR researchers (Galvez et al. 2005; Galvez and de Moya-Anegon 2006; Jacquemin and Tzoukerman 1999).

At the same time as simple conflation methods have been used in IR, not much attention has been given to heuristic linguistically motivated aids that do not even aim to cover all the inflection of the keywords but are based, for example, on the statistically most frequent word forms of the language in question. In Kettunen and Airio (2006) we showed that case form frequency based keyword generation competes quite well against the gold standard, FINTWOL lemmatizer, in best-match IR for Finnish, a highly inflectional and compound rich language. A similar but converse approach, stemming based on the statistical distribution of Hungarian noun suffixes, is reported in Tordai and de Rijke (2005). Two other types of approaches can be seen as more remotely related to our approach: Xu and Croft's (1998) idea of using corpus-based word variant statistics in stemmer creation or modification and the use of a probabilistic (and thus language independent) model for stemmer generation (Bacchin, Ferro and Melucci, 2004; Di Nunzio et al. 2004). Our method is called FCG (for Frequent Case (form) Generation).

In this paper we shall further examine our method in monolingual IR of morphologically complex languages by testing three more languages, German, Russian and Swedish, with the methods developed in Kettunen and Airio (2006). For Finnish we shall also show some new results with very short queries.

On a general level, our background motivations can be stated as follows:
The average precision and recall (P/R) of retrieval needs to be kept as high as possible without using excessively complex language technology tools; we believe that the need of large lexicon-based lemmatizers in basic monolingual IR is not as high as often thought even for a morphologically complex language.

Our research questions are following:

1) Is the FCG approach viable across languages of varying morphological complexity?

> 1a) In order of increasing complexity, what is the performance of FCG in, Swedish, German, Russian and Finnish as observed in generally available test collections?
>
> 1b) How many morphological surface forms are needed to achieve reasonable performance?
>
> 1c) How does this performance compare to doing nothing at all, stemming and lemmatization?

2) What is the effect of topic length on the performance of FCG as compared to doing nothing at all, stemming or lemmatization?

The main research question of the paper is, whether our FCG method can be shown to work with other languages that have non-trivial morphology. As the idea of the method is based on the skewed distributions of word form frequencies, it is supposed to work regardless of language in question, but verification for more than one language (Finnish) is also needed.

The performance of our new methods is compared to the state of art, usage of a lemmatizer, which is more challenging than use of raw words that has become all too common in IR (e.g., Hollink et al. 2004; Braschler and Ripplinger 2004; Mayfield and McNamee 2003; Tomlinson 2004a,b). We have argued in Kettunen, Kunttu and Järvelin (2005) that the performance gained with raw words is quite meager and variable for a morphologically rich language like Finnish, and thus the performance gains attributed to different morphological processing methods are not as great as they are thought to be. If comparisons are made, they should be made with respect to the state of the art or gold standard, not with respect to the worst possible result, as now is done many times in IR. With morphologically complex languages the best retrieval result is usually attained through a lemmatizer, such as TWOL for different languages (Koskenniemi 1996). This line of argumentation is taken in the present study.

The structure of our paper is following. First we discuss distributions of word forms in the light of linguistic corpus statistics and introduce our word form frequency based method and IR results of Kettunen and Airio (2006). After this our frequency based keyword generation method is introduced, tested and discussed using three European languages of increasing level of morphological complexity, Swedish, German, and Russian

2. Distributions of word forms

It is well known that the distributions of words and word forms are not even in texts. Some word forms occur often, some are rare. Even the distributions of different morphological categories have rates of their own, and both semantic and morphological factors play a role in distribution of word form frequencies (Baayen 1993, 2001; Manning and Schütze 1999). Karlsson (1986, 2000), e.g., shows with some semantically distinctive word types, how the case distributions of the words differ in Finnish. A word denoting a place, like *Helsinki*, has besides the dominating nominative and genitive singular forms mainly occurrences of locative cases. A person's name like *Martti* occurs mostly in nominative singular. Same sort of analysis is given by Kostić et al. (2003) for Serbian, although they seem to be hesitant about the semantic origins of the phenomenon. We shall not explore the semantic factors of case distribution any deeper, but analyze the distribution of cases on morphological level only.

In Kettunen and Airio (2006) we first sought for corpus statistics of Finnish nominal word forms. Then we verified these statistics with two independent automatic analyses of larger corpuses. Our analysis and earlier corpus statistics showed, that six cases (out of 14) constituted about 84 – 88 % of the token level occurrences of case forms for nouns – thus covering 84 – 88 % of the possible variation of about 2000 distinct inflectional forms of nouns. Our analysis also showed that the huge number of grammatical forms is mainly due to clitics and possessive endings that are almost nonexistent even in a reasonably large textual corpus (10.3 M nouns). This analysis demonstrated that, while a language may in principle be morphologically complex, in practice it is much less so.

2.1 Distribution based handling of keyword variation for IR

Our FCG (Frequent Case (Form) Generation) method and its language specific testing are simply as follows:

- For a morphologically complex enough language the distribution of different nominal case/other word forms is first studied through corpus analysis (if such results are not available for the language). The corpus used can be quite small, because variation at this level of language can be detected even from smaller corpuses. Variation in textual styles may affect slightly the results, so a style neutral corpus is the best. If style specific results are sought for, then an appropriate corpus needs to be used in word form occurrence analysis.
- After the most frequent (case) forms for the language have been found with corpus statistics, the IR results of using only these forms for noun and adjective keyword forms are tested. As a comparison best available normalization method (lemmatization or stemming) is used. The number of tested FCG processes depends on the morphological complexity of the language: more processes can be tested for a complex language, only a few for a simpler one.
- After testing, the best FCG process with respect to normalization is usually distinguished. The testing process will probably also show that more than one FCG process is giving quite good results, and thus a varying number of keyword forms can be used for different retrieval purposes, if necessary.

We have been simulating the process of keyword generation in our tests, but as word form generation programs are available for many languages, their output could be modified accordingly for real use, i.e., only the most frequent forms of generated forms would be used in search.

Based on this method, we tested four different FCGs in two different full-text collections of Finnish, TUTK (with multi-valued relevance; Sormunen, 2000) and CLEF 2003 (with binary relevance; Peters, 2003). The results of Kettunen and Airio (2006) showed that frequent case form generation works in full-text retrieval of inflected indexes in a best-match query system and competes at best well with the gold standard, lemmatization, for Finnish. Our best FCG procedures, FCG_9 and FCG_12 - with 9 and 12 variant keyword forms - achieved about 86 % of the best average precisions of FINTWOL lemmatizer in TUTK and about 90 % in CLEF 2003. We thus performed successful information retrieval of Finnish with nine and twelve variant keyword forms, which is 0.48 % and 0.64 % of the possible grammatical forms of Finnish nouns (∑ = 1872) and about 34.6 % and 46.2 % of the productive forms (∑ = 26).

One possible bottleneck of the method, too slow index search with many key forms, was also analyzed in Kettunen and Airio (2006): runtimes of the FCG queries were shown to be comparable to those of the other methods with 60 queries of the CLEF 2003 collection. Thus a hitherto unused method, frequent case form generation for morphologically complex languages, appears as a simple and effective alternative to more traditional methods like lemmatization or stemming in IR.

In Kettunen and Airio (2006) we had typical long queries made out of title and description fields of the CLEF 2003 topics. These results are replicated in Table 1.[1] For comparison, we now made also very short queries out of the title fields (mean length 2,55 words when stop words were omitted) only for the five best methods of our earlier study (plus topic words as such). Results of these runs are in Table 2.

Table 1. Finnish CLEF 2003 results, 45 title-description queries

Method	**Mean average precision**
FINTWOL, compounds split	50.8 %
Stemmed	49.8 % (-1.0)
FINTWOL, compounds not split	48.2 % (-2.6))
FCG_9	46.1 % (-4.7)
FCG_12	45.8 % (-5.0)
Inflected	31.1 % (-19.7)

Table 2. Finnish CLEF 2003 results, 45 title queries

Method	**Mean average precision**
FINTWOL, short, compounds split	42.8 %
Stemmed, short	41.3 % (-1.5)
FINTWOL, short, compounds not split	40.5 % (-2.3)
FCG_12, short	38.1 % (-4.7)
FCG_9, short	37.9 % (-4.9)
Inflected, short	22.6 % (-20.2)

[1] Results for the long queries are now recalculated for 45 queries used in the analysis. Therefore the results of Table 1 differ marginally from results of Kettunen & Airio (2006).

As can be seen from Tables 1 and 2, difference between best FCG methods and best achieved results, FINTWOL with index where compounds are split, are about 5 absolute per cent both with long and short queries. Thus the method works also well with short and realistic queries, and about 88 % of the maximal retrieval result is achieved with both nine and twelve most frequent nominal forms of the keywords.

To further analyze FGC's performance against best normalization results and worst results with inflected keywords, we did a query-by-query analysis for title-only queries. The query-by-query histograms therefore indicate how much better (upward pointing histograms) or worse (downward pointing histograms) the FCG method is compared to the baseline. In Figure 1 the best Finnish FCG results are compared to the best lemmatization results. In Figure 2 the best Finnish FCG results are shown with results of inflected keywords.

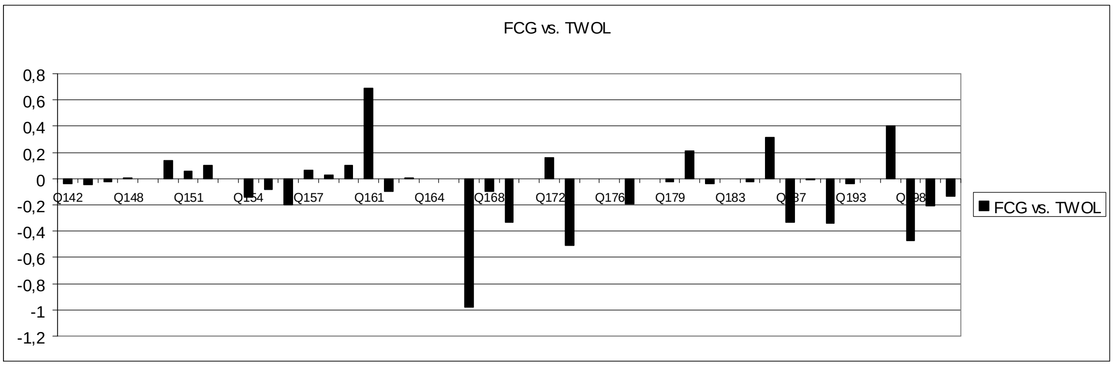

Figure 1. Query-by-query results of Finnish title-only queries. Best FCG vs. best lemmatization.

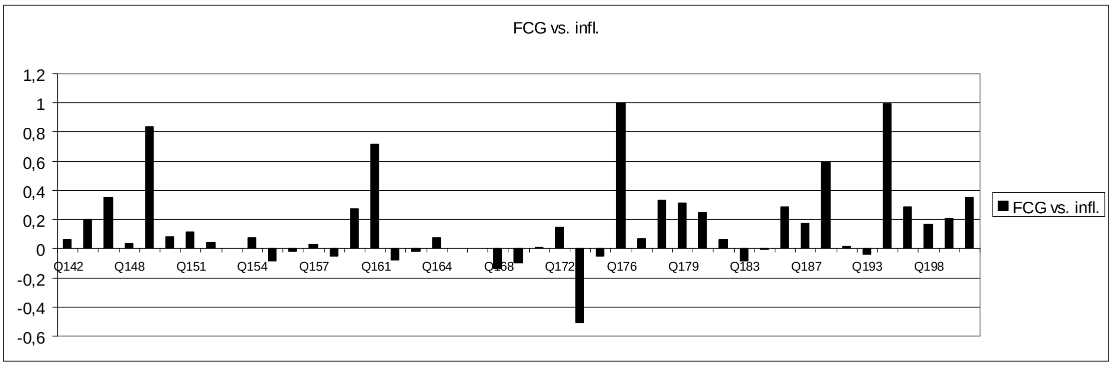

Figure 2. Query-by-query results of Finnish title-only queries. Best FCG vs. inflected queries.

When compared to the best lemmatization result, FCG was inferior in 23 queries and superior in 15 queries. There were 7 ties between lemmatization and FCG. When compared to the inflected queries, FCG won in 32 queries and remained inferior in 12 queries. Only in one query there was a tie between the methods.

In statistical significance tests FINTWOL with split compounds, Snowball and the best FCG, FCG_9, were statistically significantly better in long queries than inflected queries when Friedman test was used ($p < 0.0001$). In short queries FINTWOL with split compounds and Snowball were statistically significantly better than inflected queries ($p < 0.0001$). FCG_12 was also statistically significantly better than inflected queries ($p < 0.01$).

It is also noteworthy, that the marginal cost of doing nothing to query words (inflected) and all the tested methods is 15 – 20 absolute per cent. In practice this means that if query words in a highly inflectional language are not processed in any way, documents will be lost in searches. And as Kettunen (2006, 474) notifies, the difference between doing nothing and processing words somehow morphologically will in practice be even greater, because users will most probably use base forms of keywords in searches, even if the indexes are inflected (as, e.g., in the web).

In this study we shall test our word form frequency based method with three more European languages, Swedish, German, and Russian. They are all morphologically moderately complex, i.e. clearly much more complex than English, but also clearly much simpler than Finnish (or Hungarian, cf. Tordai and De Rijke 2005) measured in the number of possible word forms per lexeme. The chosen languages represent two major language groups of the Indo-European language family, Germanic (German and Swedish) and Slavonic (Russian), and are thus also characteristic samples for other languages in the same language groups (Comrie 1990). The languages were chosen on the basis of available IR collections and complex enough nominal morphology from the CLEF materials. From the morphological complexity point of view (cf. Kettunen et al. 2006) there would have been other and perhaps more interesting languages among the official EU languages (e.g., Estonian, Lithuanian, Latvian, Slovak, Czech and Hungarian), but either lack of available IR collections or detailed enough linguistic knowledge in the languages made inclusion of these languages impossible in this study.

3. Materials and methods

CLEF collections for all the three languages were utilized in this study. For Swedish and German we used materials of CLEF 2003 (Peters 2003). The retrieval system was InQuery (Broglio, Callan and Croft 1994). For Russian we used Russian collection of CLEF 2004 (Peters 2004) and the Lemur retrieval system (Metzler and Croft 2004; Lemur web pages). Character encoding for Russian was UTF-8. In Table 3, the number of documents and topics in each collection is shown (Airio 2006; Tomlinson 2004b).

Table 3. Swedish, German and Russian collections used in the study

Language	Collection	Collection size (docs)	Topics with relevant documents	Retrieval system in tests
Swedish	CLEF 2003	142 819	54 (out of 60)	InQuery
German	CLEF 2003	294 809	56 (out of 60)	InQuery
Russian	CLEF 2004	16 716	34 (out of 50)	Lemur

As can be seen in Table 3, the Russian collection is small. Besides the small number of documents in the collection, the number of relevant documents is also very small, only 123.

3.1 Language resources used in normalization and query generation

For normalization of the database indexes and queries, the following tools were used: SWETWOL, FINTWOL and GERTWOL lemmatizers from Lingsoft Ltd. For stemming we used Snowball stemmer's Finnish, German, Russian and Swedish versions (Snowball web page). Unfortunately there was no Russian lemmatizer available.

A *lemmatizer* analyzes inflected word forms and returns their base forms or lemmas. If an inflected word form is ambiguous, several base forms are returned. The analysis of a lemmatizer is based on a set of rules and use of a large lexicon with tens of thousands of entries[2]. TWOLs used in the study are quite typical lemmatizers that use large lexicons. *Stemming* is a many-to-one mapping where semantically related distinct word forms are reduced to identical stems either by using only affix rules or rules and lexicons. Stems that are returned by the stemmer may be linguistically motivated or only heuristic truncations of the original word forms. Snowball stemmers are typical rule-based affix removal stemmers that do not use (large) lexicons.

Our FCG method for keyword form generation was simulated for each language. The generation of the keyword forms for the FCG queries for each language was rule-based even if manual, and the following tools were used to check or generate the word forms.

For Swedish we used an electronic Swedish-Finnish dictionary Lexin and online version of SWETWOL lemmatizer. If a word form could not be verified from these, it was sought for in the Swedish web pages of the Internet or in a printed dictionary. If the word was not found in any of the sources used, it was left in the query as it originally appeared in the topic.

The German word forms for the queries were generated with Canoo net's German word form generator. Canoo's pages both analyze inflected forms and generate all the possible grammatical forms of the given base form. Also compound analysis is performed.

[2] If the word can not be analyzed, i.e. when it is misspelled or missing from the dictionary of the lemmatizer, it is marked with @ and put in a separate index of unknown words. These are sought for separately in retrieval.

Russian keywords from the CLEF 2004 topics were sought for in Multitran's web dictionary, which also gives a translation for the words and generates the different forms for the words. Russian morphological analyzer of Gelbukh and Sidorov (2003) was also utilized in checking of the forms.

3.2. Query generation and structuring

Structuring of the queries was done by using InQuery's synonymy operator, #SYN (Lemur uses also Inquery's structural operators). Query generation for lemmatization, stemming and inflected forms was automatic. Queries for the FCG test runs were formed partly manually from the topics. After automatic InQuery query structure generation, the needed case forms for query words were added with help of the electronic dictionaries and generators. Thus we simulated carefully the effects of automated rule-based frequent case form generation.[3]

As an example we can take one query from the CLEF 2003 collection. A short version of query #142 for the Sv-FCG_4 process is as follows:

> *#q142 = #sum(#syn(christo) #syn(paketerar) #syn(det) #syn(tyska tyskt) #syn(riksdagshuset riksdagshus riksdagshusen));*

The queries are thus of the form #SUM(#SYN() #SYN()…), and they are strongly structured (Kekäläinen 1999). Morphological variant forms of the keyword are treated as synonyms of the key, and InQuery treats them as instances of one key (Broglio et al. 1997). As can be seen from the query example, only nouns and adjectives of the query are expanded with variant forms, all others are left in the form they were in the original topic.

4. Morphology and morphological statistics of the three languages

We shall discuss the features of the three languages, Swedish, German, and Russian, in this chapter. In each case, we consider the morphology of the language, its nominal word form distribution for nouns and adjectives, and present the FCG processes considered

4.1 Swedish

4.1.1 Morphology of Swedish nouns

[3] If a real interactive version of the FCG system were available, users would give keywords in the manner they give them now in interactive searches, the only requirement being that keywords are given in the base form. This they should be able to do, because they are usually also able to use printed dictionaries that are based on alphabetical listings of *base forms***.**

Swedish is considered morphologically as a slightly complex language, where variation in word forms is both due to number of forms and usage of compounding (Ahlgren 2004). Also homography of Swedish word forms is very high, and this may cause problems especially for CLIR (Hedlund, Pirkola and Järvelin 2001).

Swedish nouns inflect in three categories, which all have different suffixes and affect the form of the noun: definiteness (definite, indefinite), number (singular and plural) and case (nominative and genitive). Besides these, nouns have two distinct genders, non-neuter and neuter, but this does not affect the number of different forms a word may have. When all the three grammatical categories are combined, maximally *eight* different forms of a single noun can be formed (Ahlgren 2004).

Adopting Ahlgren's example (2004, 42) of a single maximally inflected noun, the following eight forms can be formed from a base form *stad* ('city'): *stad, städer, staden, städerna, stads, städers, stadens, städernas*. With some nouns the maximal number of forms is less than eight, e.g., some nouns lack plural suffix in indefinite form as for example words of the type *målare* ('painter). This collapses the number of different forms to six, when two forms are identical with other forms. Proper names are also an exception: they do not in general inflect with respect to case and number. Swedish adjectives agree with their head noun in gender, number and definiteness.

Compounding is also a frequent phenomenon in Swedish, and its characteristic is that it will produce single (complex) words, such as *jazzmusik* ('jazz music')

4.1.2 Distributions of Swedish nominal word forms

The distribution of different noun forms of Swedish was analyzed using a SWETWOL analysis of Helsingborgs Dagblad 1994 and Göteborgs posten 1994 texts, altogether 161 336 articles (Ahlgren 2004, 61). From these SWETWOL was able to analyze 519 496 word form types, which yielded 638 012 noun interpretations including all the ambiguous analyses. When interpretations that were marked in SWETWOL's analysis with tag <SPELLING_ERROR> and some other errors were discarded, 633 058 noun interpretations were left for distribution analysis. Distributions in Table A1 in Appendix were analyzed on the basis of the resulting 633 058 forms.

From the figures of A1 it can be seen that two forms of Swedish nouns are the most important: indefinite singular nominative (base or citation form) and definite singular nominative. Together these forms make clearly over half (57.1 %) of the occurrences of all forms. If two other most frequent forms, indefinite plural nominative and definite plural nominative, are counted, these four forms together make about 81 % of the occurrences of noun forms. Forms of genitive are quite rare in the corpus, definite singular genitive being the most common form.

4.1.3 Swedish FCGs

Based on the figures in Table A1 we ended up with two Sv-FCG procedures (Swedish FCG): Sv-FCG_2 has only the two most common noun forms (indefinite and definite singular nominative); Sv-FCG_4 has besides these also indefinite and definite plural nominative. Adjectives were put in the queries in two forms **(**definite and indefinite positive**).** All keywords of other categories (verbs, adverbs etc.) were left in the form they are in the original topic. Table 4 shows the Swedish FCG procedures that were tested against SWETWOL lemmatizer and Snowball's Swedish stemmer.

Table 4. Swedish FCG procedures

Name of the procedure	**Forms in the procedure**
Sv-FCG_2	indefinite and definite singular nominative
Sv-FCG_4	Sv-FCG_2 + indefinite and definite plural nominative

4.2. German

4.2.1 Morphology of German nouns

German is considered a morphologically complex language in IR literature (cf. Braschler and Ripplinger 2004).[4] German has four distinct case forms for nouns (nominative, genitive, dative and accusative), two numbers (singular and plural) and three genera. Out of these, case and number affect the form of the noun. The German noun inflection is also slightly different from Swedish inflection in one respect: all the inflectional case and number information is not shown in the noun itself, but in the accompanying article. Thus the number of distinct noun forms in German is less than the expected eight.

Following example (*man* in German) shows a typical declination for a German noun.

Case	**Singular**	**Plural**
Nominative	der Mann	die Männer
Accusative	den Mann	die Männer
Dative	dem Mann(e)	den Männern
Genitive	des Mann(e)s	der Männer

Maximally a German noun can have four to five different forms depending on the declination class of the noun. Many noun classes have only two or three distinct forms. (cf. inflectional tables from, e.g., Deutsche Deklination 2006 or, e.g., Helbig and Buscha 1981). Thus the homography of German noun forms is high. This might be either disadvantageous or beneficial in monolingual search of our type. Compounds are also a very common phenomenon in German, and their effect on retrieval is clear (Braschler and Ripplinger 2004).

[4] Consequences of German productive word formation are many times slightly overestimated in IR. For example, Braschler and Ripplinger (2004) mention 144 different forms for the verbs, which is quite high. But they do not mention that the importance of the verbs is not that high in IR, as nouns mostly bear the meaningful content that is searched for (Baeza-Yates & Ribeiro Neto 1999, Kettunen 2006).

4.2.2 Distributions of German word forms

We analyzed the distribution of German noun and adjective forms from the data of the Tiger corpus. The corpus contains articles from the Frankfurter Rundschau 1995–97 and consists of about 900 000 word form tokens and 50 000 sentences that are part-of-speech tagged. Moreover, the corpus contains morphological and lemma information for terminal nodes (word forms). The analysis differentiates common nouns (tagged with NN) from proper nouns (tagged with NE). Attributive adjectives (i.e., adjectives determining a head noun) are also distinguished from adverbial or predicative use of adjectives.

In this corpus 178 834 common nouns are found. Occurrences of different forms of German common nouns based on this data are shown in Table A2 in Appendix. The Tiger corpus contains also 48 946 occurrences of proper nouns. Occurrences of German proper nouns based on this data are shown in Table A3.

As can be seen from the A2 and A3, German common nouns and proper nouns behave quite distinctively. For a German proper noun only nominative and dative are frequent cases and even then the names are most probably in singular. For common nouns nominative, accusative and dative are the most common cases. Singular is also the most common number for the common noun, but anyhow plural is also frequent: all the cases have about a 30 % share of plural occurrences. Although proper names have different case distributions, this should not affect as much retrieval, because the inflection of proper names is usually not shown in the word itself but in the article that precedes it. Only in genitive the proper name may differ from nominative.

4.2.3 German FCGs

The case distribution of common German nouns is slightly problematic for creation of FCG procedures. No two cases together form more than a 60 % share of the occurrences of forms. This added to the fact that many word form types in different cases are equivalent, makes it hard to choose suitable cases for De-FCGs. However, based on the distributions in A2 and A3, we made two separate FCG procedures for German, De-FCG_2 and De-FCG_4. In De-FCG_2 we had the forms of nominative and accusative in singular and plural with singular forms of dative. In De-FCG_4 forms in genitive (singular and plural) were added with plural datives. For proper names only one form was used in De-FCG_2, and the usually distinctive genitive was added to De-FCG_4.

German adjective forms were also analyzed briefly from the Tiger corpus. The corpus contained 49 076 non-comparative (positive) forms of adjectives, which were attributive (i.e. determining a head noun). Out of these 28.6 % were in nominative, 15.2 % in genitive, 29.4 % in accusative and 29.6 % in dative. About 2/3 of the forms were in singular. As the importance of adjectives is not very great in retrieval, only the most frequent five forms, singular nominative, accusative and dative and plural accusative and dative were used for both of the De-FCG procedures.

Table 5 shows the German FCG procedures that were tested against GERTWOL lemmatizer and Snowball's German stemmer.

Table 5. German FCG procedures

Name of the procedure	**Forms in the procedure**
De-FCG_2	Nominative and accusative in singular and plural, singular dative for common nouns Singular nominative for proper names Singular nominative, accusative and dative and plural accusative and dative for adjectives
De-FCG_4	De-FCG_2 + genitive in singular and plural, plural datives. Genitive for proper names Singular nominative, accusative and dative and plural accusative and dative for adjectives

4.3 Russian

4.3.1 Morphology of Russian

Russian is a Slavic language which has the most complex morphology among the languages of this study. Besides number of different possible word forms, typical to Russian morphology is also that the morphology is fusional: "thus in the declension of nouns, it is not possible to segment one inflection encoding number and another encoding case, rather these two categories are encoded by a single formative" (Comrie 1990, 337–338; cf. also Gelbukh and Sidorov 2003; Beard 1996).

Russian nouns have six distinct cases: nominative, accusative, genitive, dative, prepositional and instrumental. Nouns have four major types of declension classes for differently ending nouns. For example, a noun meaning table, *stol,* which is a masculine o-stem, is inflected as in Table 6 (example from Comrie 1990, 338, two last case names changed to current usage; cf. Beard 1996).

Table 6. Case forms of Russian nouns

Case	SG	PL
Nominative	stol	stolý
Accusative	stol	stolý
Genitive	stolá	stolóv
Dative	stolú	stolám
Prepositional	stole	stoláx
Instrumental	stolóm	stolámi

From the example we can see that there is slight overlap in the forms, in this case forms of nominative and accusative are identical in singular and plural. This happens also in other declensional types, which makes the maximum number of different forms for a noun 10−11 instead of 12 (cf. also Koval et al. 2000). Anyhow, the number of possible noun forms is greater than in either Swedish or German, and the overlap in the actual forms does not make the identification of most frequent forms as complicated as in German. In this respect Russian seems ideal for our FCG method.

4.3.2 Distributions of Russian word forms

The distributions of different noun and adjective forms of Russian were obtained from the Russian National Corpus. Statistics in Tables A4 and A5 in Appendix are based on a 5 million word hand-tagged sub-corpus.

As we can see from the Russian data, three cases, nominative, genitive and accusative, are the most frequent ones. If singular and plural are joined, occurrences of these three cases form 75.7 % of all the word forms of nouns. Besides these, prepositional and instrumental are almost as common, instrumental being slightly more frequent in singular but prepositional in plural. Overall there also seems to be a quite big difference between occurences of singular and plural forms. Out of all 1.3 M noun forms 77 % are in singular, and only 23 % in plural. Adjectives have almost the same kind of distribution, nominative, genitive and accusative being the most frequent forms. Only the frequency of instructive and locative forms is slightly different, instructive being more common for adjectives.

4.3.3 Russian FCG procedures

Based on the distribution data we formed three FCG procedures for Russian shown in Table 7.

Table 7. Russian FCG procedures – nouns and adjectives in same forms

Name of the procedure	Cases included
Ru-FCG_3	Nominative, genitive and accusative, only singular forms
Ru-FCG_6	Nominative, genitive and accusative, singular and plural forms
Ru-FCG_8	Nominative, genitive, accusative and instrumental, singular and plural forms

5. Results

Results of the tests for the three new tested languages are shown in this section. For all the languages both long and short queries were tested.

5.1. Swedish results

Ahlgren (2004) is evidently the most thorough discussion of full-text information retrieval of Swedish (cf. also Ahlgren and Kekäläinen 2007). His research settings include different types of keywords (inflected, i.e. non-processed keywords, truncated keywords, keyword lemmas) and different types of indexes (inflected, lemmas, lemmas with compound splitting, and lemmas and compound splitting with compound elimination principle) (Ahlgren 2004, 74). His results show, perhaps slightly astonishingly, that keyword truncation search in an inflected index is the best method in the collection used with a small margin to both lemma-based searches using different types of split compound indexes (Ahlgren 2004, 102). Airio (2006) got her best monolingual Swedish results for CLEF 2003 collection by using keyword lemmas in an index, where compounds were split. Hollink et al. (2004) got their best results for Swedish with CLEF 2002 collection using stemming combined with compound splitting.

On the basis of this, and particularly based on the success of truncated keywords in Ahlgren (2004), it should be expected, that Sv-FCGs would work reasonably well. Based on the distribution of the forms in corpus, it may be expected that at least Sv-FCG_4 should perform quite well in retrieval runs compared to full coverage morphological analysis. Sv-FCG_2 may be too crude a procedure, but it should also outperform usage of plain words.

Results of the Swedish runs for long queries (average length 15.62 words with stop words) consisting of the title and description fields of the topics are shown in Table 8.

Table 8. Results of the 54 Swedish title-description queries

Method	Mean average precision
SWETWOL, compounds split [5]	38.8 %
Sv-FCG_4	35.2 % (-3.6)
Sv-FCG_2	33.7 % (-5.1)
Stemmed	33.5 % (-5.3)
Inflected	32.1 % (-6.7)
SWETWOL, compounds not split	31.4 % (-7.4)

The results of long queries of Swedish queries do not show very great differences between different keyword processing methods. The margin between non-processed keywords and best normalization result is 6.7 %. The best Sv-FCG performs 3.6 % below SWETWOL using split compound index. Sv-FCG_2 and Snowball Swedish stemmer perform almost the same level. Their difference to non-processed keywords is only 1.4 – 1.6 %. When compounds are not split in the index, SWETWOL is performing worst, slightly below non-processed query words.

We also ran very short queries made out of the title fields of topics (average length 3.17 words with stop words). Results of the Swedish runs for very short queries are shown in Table 9.

Table 9. Results of the 54 Swedish title queries

Method	Mean average precision
SWETWOL, compounds split	32.6 %
Sv-FCG_4	30.6 % (-2.0)
Sv-FCG_2	29.1 % (-3.5)
Stemmed	28.5 % (-4.1)
SWETWOL, compounds not split	26.3 % (-6.3)
Inflected	24.0 % (-8.6)

Very short queries behave almost in the same way as long queries. The margin between non-processed keywords and best normalization result is slightly larger, 8.6 %. Both Sv-FCGs outperform stemming and SWETWOL without compound splitting. All the keyword processing methods are now also clearly better than non-processing. This is

[5] Compounds are split in the index of the query system, but not in the queries, which has been tested to give best results. In practice this means that e.g. *jazzmusik* will be in the index as a whole, and also as *jazz* and *musik*, all the three words pointing to the same document location. This concerns Finnish, Swedish and German indexes and queries.

natural with short queries, which do not offer as many access points to documents as long queries.

In Figure 3 the best Swedish FCG results are compared query-by-query to the best lemmatization results for title-only queries. In Figure 4 the best Swedish FCG results are compared query-by-query to results of inflected keywords for title-only queries.

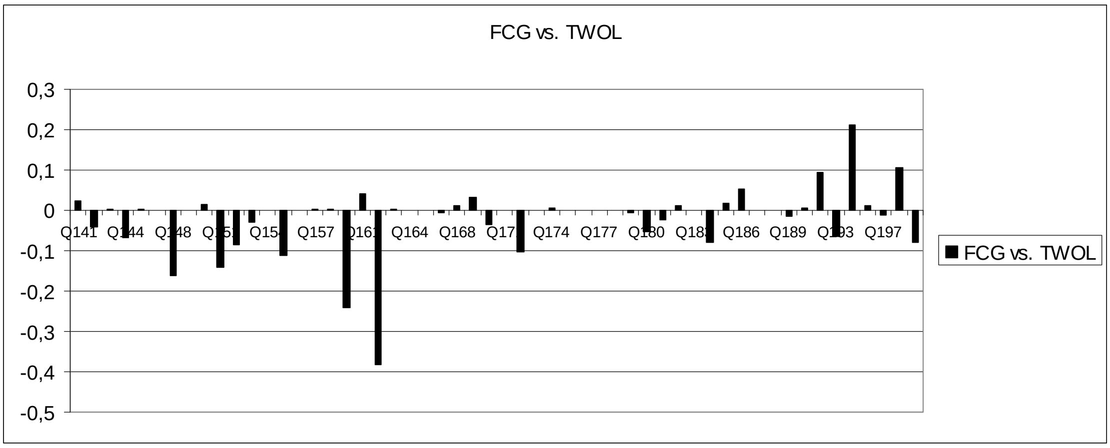

Figure 3. Query-by-query results of Swedish title-only queries. Best FCG vs. best lemmatization.

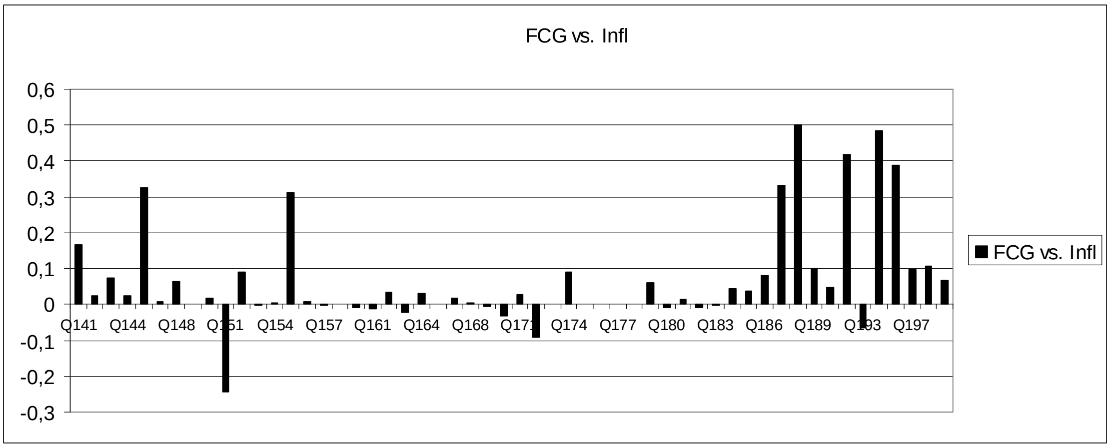

Figure 4. Query-by-query results of Swedish title-only queries. Best FCG vs. inflected queries.

When compared to the best lemmatization result, FCG was inferior in 21 queries and superior in 23 queries. 10 queries were ties between lemmatization and FCG. When compared to inflected queries, FCG was superior in 33 queries and inferior in 13 queries. There were 8 ties between the methods.

The statistical testing of the differences between methods used in the present study was done using the Friedman test (original Friedman test, cf. Siegel and Castellan, 1988, modifications used in here in Conover, 1980). The main reason for this was that multiple methods were compared to each other and that parametric test are, strictly taken, not applicable because the data do not follow their assumptions on distributions. For example, Hull (1993), Kekäläinen (1999) and Kraaij (2005), state that in such a context, the Friedman test is appropriate.

The Friedman two-way analysis of variance by ranks is a generalization of the parametric sign test. Thus it offers a non-parametric alternative for comparing more than two related samples (Hull 1993).The basic principle of the Friedman test is to first calculate, whether there are significant differences between the evaluated methods overall. If such differences are found, a pair-wise comparison between different methods is done to show which methods differ significantly from each other.

For Swedish and German tests were made between the two TWOLs, Snowball, best FCG procedure and inflected query words.

Results of the statistical significance tests for Swedish long and short queries are shown in Table 9a.

Table 9a. Statistically significant differences between Swedish methods

Method, long queries	**Statistical significance *)**
1. SWETWOL, compounds split	*3 >>4
2. Sv-FCG_4	----
3. Stemmed	----
4. Inflected	> 5
5. SWETWOL, compounds not split	----
Method, short queries	
6. SWETWOL, compounds split	>> 9
7. Sv-FCG_4	> 9
8. Stemmed	>> 9
9. Inflected	----
10. SWETWOL, compounds not split	----

*) Friedman test: >> = p < 0.01 > = p < 0.02 * = statistically almost significant, p < 0.05

5.2 German results

Problems of information retrieval in German are quite well known. The morphology of German is quite complex and especially use of compounding is common in the language.

Braschler and Ripplinger (2004) evaluate several different stemming approaches and effect of compound splitting on German IR. They found that stemming together with compound splitting is the most effective approach for German. In long queries the improvements were smaller than in very short queries. Results of Airio (2006) support the same conclusion: even the use of a lemmatizer does not increase search results much, if the compounds are not split in the index of the query system. Airio also shows that the use of a simple stemmer (Snowball stemmer for German) can be as effective as the use of a lemmatizer and compound splitting in the index.

Results of our German runs for long queries (average length 17.25 words with stop words) consisting of the title and description fields of the topics are shown in Table 10.

Table 10. Results of the 56 German title-description queries

Method	**Mean average precision**
GerTWOL, compounds split	39.7 %
Stemmed	39.1 % (-0.6)
De-FCG_4	38.0 % (-1.7)
De-FCG_2	36.8 % (-1.9)
Inflected	35.9 % (-3.8)
GerTWOL, compounds not split	35.1 % (-4.6)

The results of long German queries do not show very great differences between different methods. The margin between non-processed keywords and best normalization result is only 3.8 %. Stemming performs almost as well as GERTWOL using split compound index. The best De-FCG performs 1.7 % below GERTWOL using split compound index. De-FCG_2 outperforms also GERTWOL without split compound index. It is also noteworthy that GERTWOL with compounds in the index performs slightly worse than non-processing.

As can be seen from the results, differences between different methods are not great. One of the main explanatory reasons for this most probably is German inflectional homography: many grammatical nominal forms have the same surface form, and the precise grammatical case can be distinguished only from the use of article or context. Thus one keyword form will often hit several occurrences of different grammatical forms in the index.

Results of the German runs for very short queries (average length 3.15 words with stop words) consisting of only the title fields of the topics are shown in Table 11.

Table 11. Results of the 56 German title queries

Method	**Mean average**
GerTWOL, compounds split	29.6 %
Stemmed	30.9 % (+1.3)
De-FCG_4	29.9 % (+0.3)
De-FCG_2	29.0 % (-0.6)
GerTWOL, compounds not split	28.1 % (-1.5)
Inflected	25.4 % (-4.2)

Very short German queries show the same overall performance as long queries, but differences are smaller. This time the Snowball stemmer performs the best with a 1.3 % margin to GERTWOL using split compound index. De-FCG_4 is also slightly better than GERTWOL, and De-FCG_2 outperforms again GERTWOL without compound splitting. Non-processed queries perform now worst, and the margin of non-processing to the best performing system, Snowball, is 5.5 %. The margin of non-processing to the worst performing normalization is 2.7 %.

In Figure 5 the best German FCG results are shown query-by-query with the best stemming results for title-only queries. In Figure 6 the best German FCG results are shown query-by-query with results of inflected keywords for title-only queries.

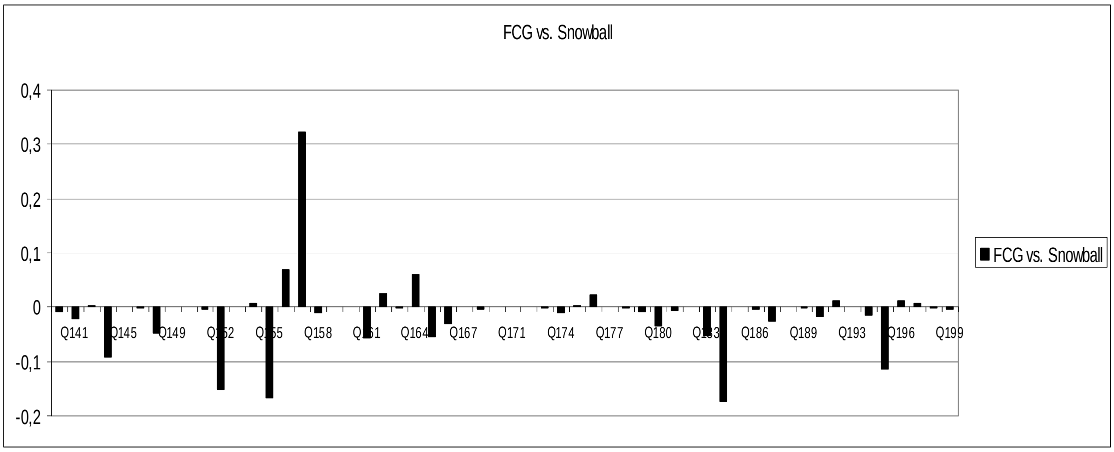

Figure 5. Query-by-query results of German title-only queries. Best FCG vs. best stemming.

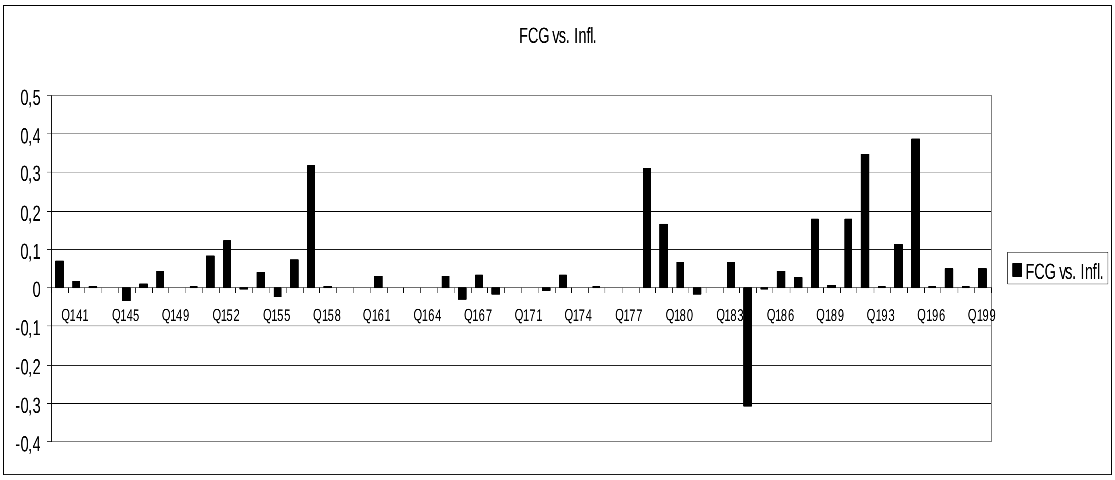

Figure 6. Query-by-query results of German title-only queries. Best FCG vs. inflected queries.

When compared to the best stemming result, FCG is inferior in 30 queries and superior in 18 queries. In 8 queries there were ties between stemming and FCG. When compared to inflected queries, FCG was superior in 36 queries and inferior in 13 queries. There were 7 ties between the methods.

Results of the statistical significance tests for German long and short queries are shown in Table 11a.

Table 11a. Statistically significant differences between German methods

Method, long queries	Statistical significance*)
1. GerTWOL, compounds split	>> 4 >* 5
2. Stemmed	>> 4 > 5
3. De-FCG_4	* 4
4. Inflected	------
5. GerTWOL, compounds not split	------
Method, short queries	
6. GerTWOL, compounds split	>* 9
7. Stemmed	> 9 * 10
8. De-FCG_4	>* 9
9. Inflected	------
10. GerTWOL, compounds not split	* 9

) Friedman test: > = $p < 0.01$ >> = $p < 0.001$ > = $p < 0.02$
* = statistically almost significant, $p < 0.05$

5.3 Russian results

Information retrieval experiments in Russian have not been reported much outside Russia. CLEF campaign introduced a small Russian collection of 16 716 articles (from Izvestia 1995) in 2003, and since then there have been a few papers concerning retrieval of Russian in CLEF. Petrasi, Perelman and Gey (2003) give a baseline approach in domain-specific Russian retrieval. Tomlinson (2004a, 2004b) shows performance of a lexical stemmer for four and nine languages, including Russian. Gey (2005) gives a baseline approach in domain-specific Russian retrieval, and Gey (2004) introduces cross-language IR with Russian documents as target.

Contrary to German and Swedish tests, our Russian tests were done using the Lemur query system, because it was able to handle documents in UTF-8 character encoding. Lemur combines an inference network retrieval model with language models, which are thought to give more sound estimates for word probabilities in documents. (Metzler and Croft 2004; Grossman and Frieder 2004). One of the key benefits of the approach is, that "the resulting model allows structured queries to be evaluated using natural language estimates" (Metzler and Crof 2004).

Russian CLEF 2004 collection has 34 topics that have relevant documents (out of 50 topics). The collection is problematic in the way that it is both small and contains only 123 relevant documents.

Table 12 presents results of the Russian runs with long queries. Long queries were made out of the topics without omitting any stop words, and the queries had 16.7 words on average (title + description fields, with stopwords). Russian results show the mean average precision and also the number of retrieved documents in top-1000 of the results. The language model smoothing method used in the runs was Dirichlet which seemed to give best results (cf. Grossman and Frieder 2004, 52–56). No pseudo-relevance feedback was used.

Table 12. Results of 34 Russian title-description queries

Method	**Mean average precision**	**Number of relevant documents returned (out of 123). Cut-off value 1000 documents.**
Snowball Ru	34.7 %	90
Ru-FCG_3	32.7 % (-2.0)	76
Inflected	29.8 % (-4.9)	78
Ru-FCG_6	29.2 % (-5.5)	88
Ru-FCG_8	28.9 % (-5.8)	95

Table 13. Results of 34 Russian title queries

Method	**Mean average precision**	**Number of relevant documents returned (out of 123). Cut-off value 1000 documents.**
Ru-FCG_6	32.0 %	84
Ru-FCG_8	31.7 % (-0.3)	86
Ru-FCG_3	31.2 % (-0.8)	78
Snowball Ru	27.2 % (-4.8)	81
Inflected	25.1 % (-6.9)	67

Results for Russian very short queries are shown in Table 13. Mean length of the title only queries was 3.18 words (with stopwords).

In Figure 7 the best Russian FCG results are compared query-by-query to the best stemming results for title-only queries. In Figure 8 the best Russian FCG results are compared query-by-query to results of inflected keywords for title-only queries.

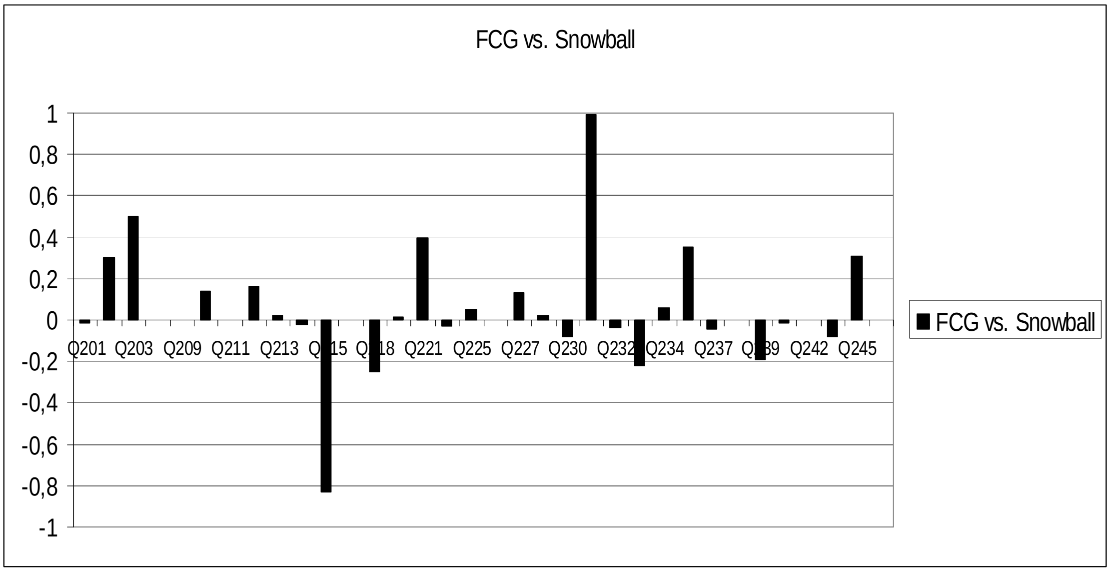

Figure 7. Query-by-query results of Russian title-only queries. Best FCG vs. stemmed queries.

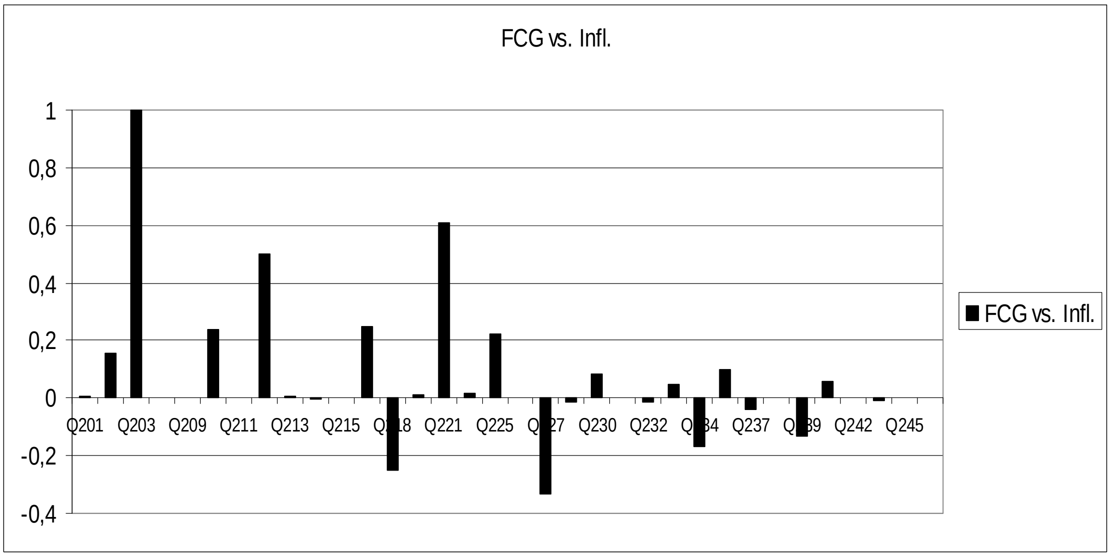

Figure 8. Query-by-query results of Russian title-only queries. Best FCG vs. inflected queries.

When compared to the best stemming result, FCG was inferior in 12 queries and superior in 14 queries. In all, 8 queries were ties between stemming and FCG. When compared to inflected queries, FCG was superior in 15 queries and inferior in 9 queries. There were 10 ties between the methods.

Statistical significance tests for the Russian data using the Friedman test showed no statistically significant differences between different methods either in long or short queries.

As can be seen from Tables 12 and 13, Russian long and short queries seem to behave differently. While Russian long queries get their best performance with Snowball stemmer, the stemmer is in short queries worse than all the Ru-FCGs and only slightly better than non-processed queries. With long queries Ru-FCG_3 is the best method against all the expectancy that more word forms should improve performance. Anyhow, the recall of both long and short queries is getting better when more word forms are added to the query in different Ru-FCGs. Even if Ru-FCG_8 is getting lower precision than Ru-FCG_3, it is able to find 19 more documents in top-1000. In addition, we note that the FCG approach is at least as good as, and statistically not different from, stemming.

6. Discussion

Our research questions for this study were formulated in the introduction as testing of the suitability of the FCG method with new languages against lemmatization and stemming with both long and short queries. We had earlier shown with Finnish that the use of only the most frequent noun and adjective forms worked well, when compared to the best available morphological method, usage of a lemmatizer. We now tested our method with three more languages, Swedish, German, and Russian, which all have a fair degree of morphological variation. For all of the languages we tested both normal laboratory type long queries and more realistic very short queries taken out of the title field of the topics.

Our Swedish results showed quite clearly that the FCG method works for Swedish in both long and short queries. In short queries differences between all methods are smallest, but the margin between non-processing and the best method also increases, which emphasizes the meaning of some sort of keyword processing. Lemmatization with compound splitting is the best method in both long and short queries.

Our German results showed that the method works for German too, although the overlap of inflected noun forms slightly disturbs results. The margin between non-processing and the best method is smaller than in Swedish, which is obviously due to inflectional homography. However, the differences of FCGS from the gold standards were statistically insignificant.

Our Russian results are partly counterintuitive. With both long and very short queries recall rises steadily when more case forms are put into the query. Anyhow, the mean precision of long queries does not get any better when forms are added, but rather decreases. The best mean average precision with long queries is gained with process Ru-FCG_3 with the singular forms only. But the inflected queries, where query words are taken as such from the topics, are the third best method in terms of mean average precision with long queries. Overall it seems that short Russian queries show some advantage for FCGs, but as the collection is small and has very few relevant documents, the interpretation of the Russian results remains inconclusive.

In very short Russian queries the difference between doing nothing (inflected queries) and the use of a different number of case forms is quite clear. Anyhow, the differences between different case form procedures in terms of mean average precision are very small, and only recall clearly improves when more forms are used.

The main reason for using stemming, lemmatization or any kind of morphological processing with IR is improvement in precision and recall of searches. Although the gains of morphological processing are varying, they are real. The usual way to estimate the performance gains is relative percentage improvement of mean average precisions between different methods. For comparison purposes of methods a slightly different point of view could also be used: the difference between doing nothing for the query words and the best mean average precision shows the need of morphological processing for the language in question. The bigger the discrepancy between these figures, the bigger the need to do something for the keywords.

In Figures 9, 10, 11 and 12 we show P-R curves of Finnish, Swedish, German and Russian short queries for the best normalization method, best FCG method and no processing at all. Short queries are shown here, because they represent more realistically real user searches. As can be seen from figures and Tables 2, 9, 11 and 13, the largest difference between non-processing and best normalization method is in Finnish (20.4 %) and smallest in Swedish (4.1 %). German and Russian have slightly greater differences than Swedish, 5.7 % and 6.9 %, respectively. Figures also show that the FCG method gives clear gains for Finnish and smaller gains for German, Swedish and Russian. However, for three languages FCG works well in comparison to lemmatization; for Finnish 88 % of the performance of lemmatization is achieved and 95 % for Swedish and German. The graphs also show that the FCG method pushes close to normalization even when the gap between normalization and non-processing is narrow.

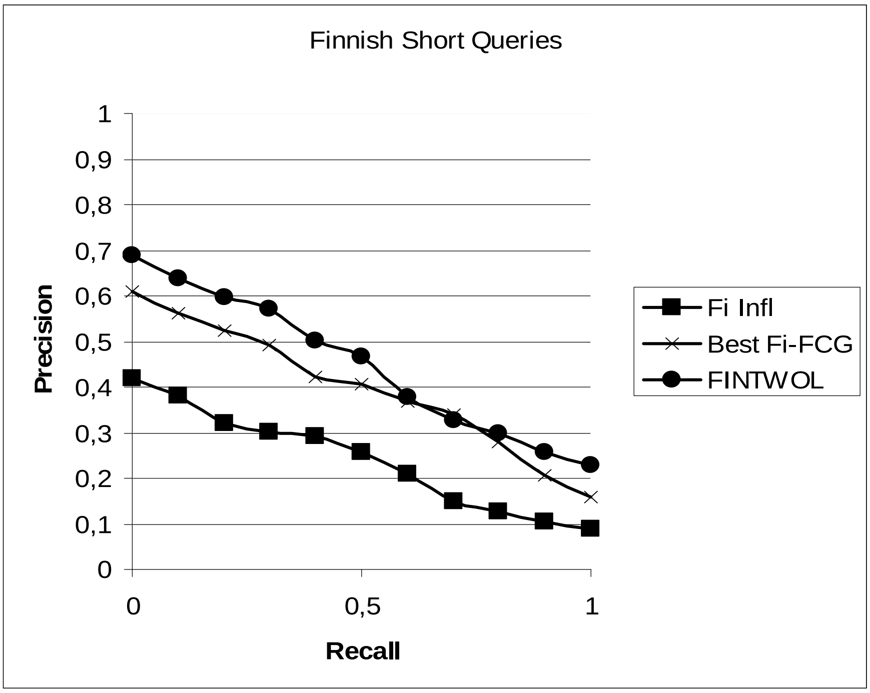

Figure 9. P/R curves for Finnish short queries: precision by eleven recall levels 0.0 -1.0

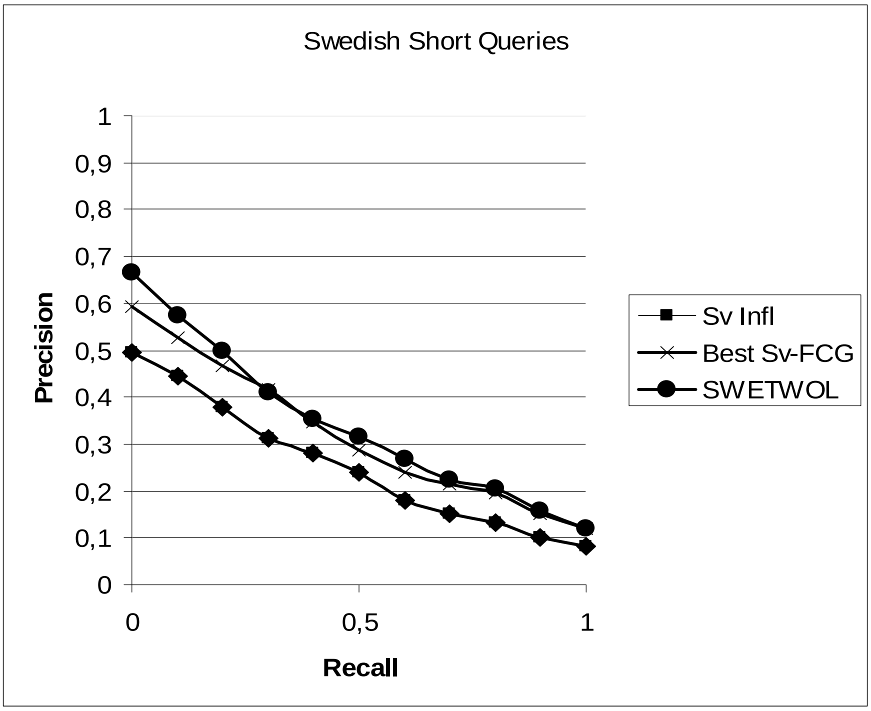

Figure 10. P/R curves for Swedish short queries: precision by eleven recall levels 0.0 - 1.0

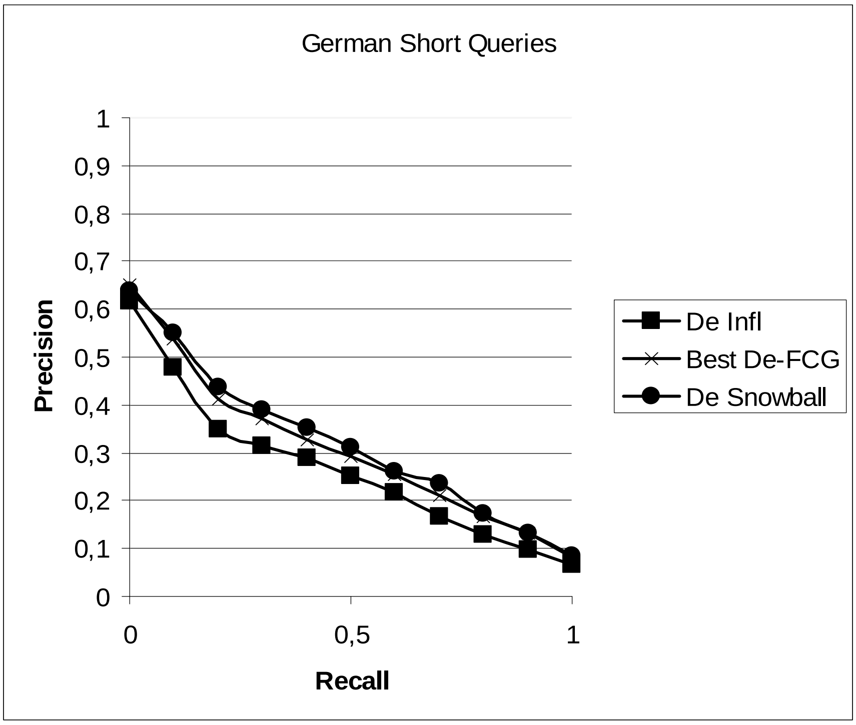

Figure 11. P/R curves for German short queries: precision by eleven recall levels 0.0 - 1.0

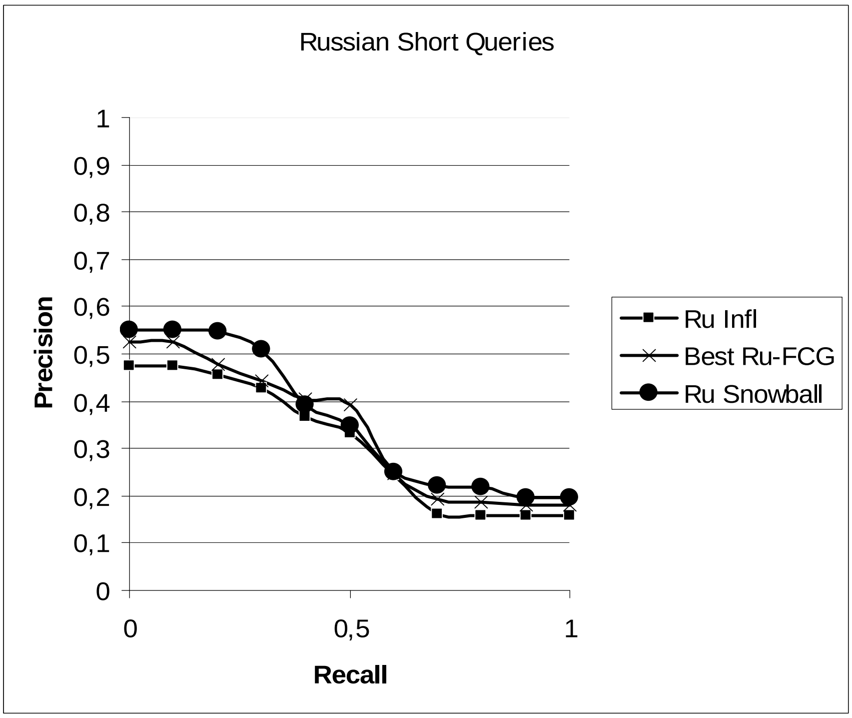

Figure 12. P/R curves for Russian short queries: precision by eleven recall levels 0.0 - 1.0

It may be foreseen that a major application area for the FCG approach is Web searching. The present state of language specific search capabilities of general search engines, such as Google, Alltheweb or Altavista, does not seem satisfying from the user point of view. Very few search engines seem to offer e.g. stemming, and search term truncation has been omitted almost totally (Search Engine Showdown 2006). The status of language specific search capabilities of general search engines thus seems poor. Bar-Ilan and Gutman (2005) report their findings for four different languages (French, Hebrew, Hungarian and Russian, tests made in November 2002) with national and general search engines. From their results it can be seen that national web services (such as Yandex in Russian, Origo-Vizsla in Hungarian and Morfix for Hebrew) take into account the requirements of each particular language and their search results are far better than those of general search engines with the language in question. As the web is constantly becoming more multilingual (Bar-Ilan and Gutman 2005; Greffenstette and Nioche 2000), it would also be desirable if the most popular search tools of the web were more sensitive to the language-specific requirements. Otherwise the huge information potential of the non-English web cannot be effectively utilized.

The method we have presented in this paper and Kettunen and Airio (2006) provides one effective solution to the problem of web searches in various languages. So far we have shown that it competes well with other morphological programs in languages of varying morphological complexity. The basic idea of the method is easily adaptable to other languages and testing of the effects of FCG style of search can be implemented relatively easily with the present state of language technology tools and search engines.

We have generated in each FCG case mechanically all inflectional forms of nouns, no matter what their semantic category (e.g. a person’s name, a word denoting to a place etc.). Some mileage may be gained by partitioning the generation to different case sets by the semantic noun category because the case distributions vary by the noun category. The mileage is however gained only through the cost of identifying the noun categories and therefore we do not further consider this possibility in the present paper.

Why, then, use this kind of approach when full morphological analysis programs are available? There are several reasons for this.

First, the generation approach works with inflected indexes of search systems and no base form processing is needed for the index. This is mainly of practical value, but an important issue: as web indexes especially are very large, separate base form runs for them would take a great deal of time. As indexes also need constant updating, making base form indexes does not sound like a very good option. Searching the index with a few most frequent inflected forms of keywords should not take too much time, when the usual web search consists of only one to three keywords regardless of the language (Jansen, Spink and Saracevic 2000; Jansen and Spink 2005).

Secondly, our approach, generation of only the most frequent word forms, is simple and could be easier to implement, if (usually) commercial morphological analyzers are not available for IR in a specific language. Earlier, in Kettunen, Kunttu & Järvelin (2005), we tried out inflectional stem generation based retrieval. Performance of the runs was good, but the queries constructed with this method (by harvesting the inflectional index with inflectional stems) were large, up to circa 2−54 * the original query length in query terms.

In Kettunen (2006) this problem was remedied by using regular expressions to enhance inflectional stems to make queries smaller. This method worked better with respect to query size, but P/R declined somewhat. Our FCG method, then, can be seen as an optimized compromise out of these earlier generative efforts.

Thirdly, while word based IR is quite effective, it requires handling word form variation in some way. Although full-scale morphological programs perform well, as Galvez et al. (2005) and Galvez and de Moya-Anegón (2006) argue, they may be unnecessarily complex and resource consuming for this purpose.

Fourthly the generation approach is not as dependent on large lexicons as are full-scale morphological analyzers, because for many languages use of lexicon in generation is not necessary. The main advantage of not using large lexicons is that out-of-vocabulary words do not affect retrieval results, as they evidently do with lemmatizers.

For the languages tested so far with the FCG approach, realistically long web-style searches would mean longer searches than with one form. In Table 14 we present the mean number of word forms per lexeme that are maximally generated for our short queries for each language's best FCG procedure. From this the number of required search forms can be realistically approximated.

Table 14. Mean number of generated word forms per lexeme for each language in short queries, stop words not included.

Language	**Forms/lexeme**
Finnish	12.27 (FCG_12) 9.35 (FCG_9)
Swedish	3.29 (Sv-FCG_4)
German	2.98 (De-FCG_4)
Russian	5.34 (Ru-FCG_8) 3.80 (Ru-FCG_6)

As we can see in the table, the figures for German and Swedish are not prohibitively high. In a typical one to three word web search, these figures would mean about 3–10 keyword forms for German and Swedish. For Russian searches the number of generations for Ru_FCG_8 would mean already about 5–16 keyword forms, which is rather high. Ru-FCG_6 would generate 4–12 keyword forms. For Finnish the number of keyword forms, 9–36, might border on the impractical, but good index packaging and retrieval algorithms might make even this possible. A smaller number of generated keyword forms, six, (cf. Kettunen and Airio 2006) could be enough for Finnish. In Kettunen and Airio (2006) we also evaluated, how the increasing number of keyword variants affects query runtime.

When the maximal 12 Finnish keyword variants were used in long queries, the increase in mean CPU time was only about 20 % in comparison to minimum, three forms or plain unprocessed keywords. An explanation is that often the different inflectional forms of a keyword reside on the same index page, which means that additional disk accesses are not required so often. This shows that processing time of queries does not increase prohibitively at least in a laboratory retrieval environment of moderate size when number of keyword variants is increased 3−12 fold. The consequences of the FCG approach for a real multi-user search system's search times are matter of implementation and can not be evaluated here.

7. Conclusions

Morphological normalization is needed in IR because morphological variation needs to be accounted for and several natural languages are morphologically non-trivial. Several languages have a fairly complex or very complex morphology in the sense, that each nominal base form may have 8–26 productive variant forms (and even a few thousand grammatical forms). There are two basic and popular approaches to morphological normalization: lemmatization and stemming. Lemmatization is effective but often requires expensive resources. Stemming is also effective in most contexts, generally almost as good as lemmatization and typically much less expensive; besides it also has a query expansion effect. However, in both approaches the idea is to turn many inflectional forms to a single lemma or stem both in the database index and in queries. This means additional database normalization for indexes.

In this paper we took an opposite approach, called FCG (for Frequent Case (form) Generation): the database index is left un-normalized and the queries are enriched to cover for surface form variation of keywords. We have shown that word form generation of 3–9 most frequent cases or forms is sufficient. Therefore the potential penalty of the approach in processing time remains negligible. By only covering such a negligible number of surface forms even in morphologically complex languages one may arrive at a performance that is just as good as or better than what stemming or lemmatization provide. It is also noteworthy, that for morphologically less complex languages, as e.g. Romance languages mostly are, this means that all the varying noun forms can be generated for retrieval.

The method has been tested with four languages, Finnish, German, Russian and Swedish, in IR laboratory settings using newspaper article collections of various sizes and queries of varying length. The results for Finnish, German and Swedish are clearly positive. In Finnish the best FCG methods achieve about 86–90 % of the best lemmatization results. In Swedish the best FCG methods achieve about 90–93 % of the best lemmatization results. In German the best FCG methods achieve about 96–97 % of the best lemmatization or stemming results. The best FCG method in these languages was never significantly worse than the best lemmatization or stemming method. Furthermore, in these three languages at least one FCG procedure was significantly better than doing nothing to keyword variation. Russian results are ambivalent, and they should be retested in a better collection, when such becomes available. On the basis of these findings and common knowledge about word form distributions in texts of natural languages, it is to be expected, that the method will work for other languages of equal morphological complexity as well. Nevertheless the findings of this paper should be tested further with more languages and in different query environments to make sounder conclusions about the proposed method's applicability in IR Applications include in particular Web IR in languages poor in morphological resources (sometimes called "low density languages"). Also the multilinguality of a web index (Rasmussen 2003) can be dealt with the approach.

Acknowledgements

Ph.D. Mihail Mihailov (Department of Translation Studies, University of Tampere) has helped with details of Russian word formation. Ph. D. Grigori Sidorov (Center for Computing Research, Mexico) provided a Russian inflectional generator for use. Ph. D. Harald Lüngen (Justus-Liebig Universität, Giessen, FB 05 - Applied and Computational Linguistics) gave helpful comments on German inflection. We are also grateful to the FIRE research group for helpful comments.

The InQuery search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst.

The Lemur query system is available from http://www.lemurproject.org/. It is "a collaboration between the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University".

The Snowball stemmers for Finnish, German, Russian and Swedish are available from the Snowball web site, http://snowball.tartarus.org/.

This research was supported, in part, by the Academy of Finland Grant No. 204978.

References

Ahlgren, P (2004), The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database. Department of Library and Information Science/Swedish School of Library and Information Science. University college of Borås/Göteborg University.

Ahlgren, P and Kekäläinen, J (2007) Indexing strategies for Swedish full text retrieval under different user scenarios. Information Processing and Management 43: 81–102.

Airio, E (2006) Word Normalization and decompounding in mono- and bilingual IR. Information Retrieval 9: 249–271.

Baayen, R H (1993) Statistical Models for Word Frequency Distribution. Computers and the Humanities 26: 347–363.

Baayen, R H (2001) Word Frequency Distributions. Kluwer Academic Publishers, Dordrecht Boston London.

Bacchin, M, Ferro, N and Melucci, M. (2004) A probabilistic model for stemmer generation. Information Processing and Management 41(1), 121 – 137.

Baeza-Yates, R and Ribeiro-Neto B (1999), Modern Information Retrieval. Addison Wesley, USA.

Bar-Ilan, J and Gutman, T (2005) How do search engines respond to some non-English queries? Journal of Information Science 31: 13–28.

Beard, R (1996) An interactive on-line Russian reference grammar. http://www.alphadictionary.com/rusgrammar/ (visited September 25th, 2006).

Braschler, M and Ripplinger, B (2004) How Effective is Stemming and Decompounding for German Text Retrieval? Information Retrieval 7: 291–316.

Broglio, J, Callan, J, Croft, WB 1994. INQUERY System Overview. In: Proceedings of the TIPSTER text program (Phase I). San Francisco, CA: Morgan Kaufmann Publishers.

Canoonet. Free Online German language resources, www.canoo.net (visited August 2006).

Comrie, B (1990) Russian. In: Comrie, B (ed.), The World’s Major Languages. Oxford University Press, New York, pp. 329–347.

Conover, W J (1980) Practical Nonparametric Statistics. 2nd edition. John Wiley and Sons, New York.

Deutsche Deklination, http://de.wikipedia.org/wiki/Deutsche_Deklination (visited June 8th, 2006).

Di Nunzio, G M, Ferro, N, Melucci, M, and Orio, N (2004) Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In: Peters, C et al. (eds.), Comparative Evaluation of Multilingual Information Access Systems. LNCS #3237, Springer-Verlag, Berlin, 2004, 220–235.

Galvez, C and de Moya-Anegón, F (2006) An evaluation of conflation accuracy using finite-state transducers. Journal of Documentation 62: 328–349.

Galvez, C, de Moya-Anegón, F and Solana, V H (2005) Term Conflation Methods in Information Retrieval. Non-linguistic and Linguistic Approaches. Journal of Documentation 61: 520–547.

Gelbukh, A and Sidorov, G (2003) Approach to Construction of Automatic Morphological Analysis Systems for Inflective Languages with Little Effort. In: Computational Linguistics and Intelligent Text Processing (CICLing-2003, Mexico City). Lecture Notes in Computer Science N 2588, Springer-Verlag, pp. 215–220.

Gey, F (2004) Searching a Russian Document Collection using English, Chinese and Japanese Queries. Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK. http://clef.isti.cnr.it/2004/working_notes/WorkingNotes2004/20a.pdf (visited October 10th, 2006).

Gey, F (2005) Domain-Specific Russian Retrieval: A Baseline Approach. Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/gey05.pdf (visited October 10th, 2006).

Grefenstette, G and Nioche, J (2000) Estimation of English and non-English language Use on the Web. http://arxiv.org/ftp/cs/papers/0006/0006032.pdf (visited 28th October, 2006).

Grossman, D A and Frieder, O (2004) Information Retrieval. Algorithms and Heuristics. Second edition. Springer, Netherlands.

Helbig, G and Buscha, J (1981) Deutsche Grammatik. 7. unveränderte Auflage. VEB Verlag Enzyklopädie, Leipzig.

Hedlund, T, Pirkola, A and Järvelin, K (2001) Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Information Processing and Management 37: 147–161.

Hollink V, Kamps J, Monz C and de Rijke, M (2004) Monolingual document retrieval for European languages. Information Retrieval 7: 33−52.

Hull, D (1993) Using statistical testing in the evaluation of retrieval experiments. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 329–338.

Jansen B and Spink, A (2005) An analysis of Web searching by European Alltheweb.com users. Information Processing and Management 41: 361−381.

Jansen, B, Spink, A and Sarasevic, T (2000) Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. Information Processing & Management 36: 207–227

Jacquemin, C and Tzoukerman, E (1999) NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax. In: Strzralkowski, T (ed.): Natural Language Information Retrieval. Kluwer Academic Publishers, Dordrecht Boston London, pp. 25–74.

Karlsson, F (1986) Frequency Considerations in Morphology. Zeitsschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 39: 19–28.

Karlsson, F (2000) Defectivity. In: Booij G et al. (eds.): Morphology. An International Handbook on Inflection and Word-Formation. Volume 1. Walter de Gruyter, Berlin, pp. 647–654.

Kekäläinen, J (1999) The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Acta Universitatis Tamperensis 678.

Kettunen, K (2006) Developing and automatic linguistic truncation operator for best-match retrieval of Finnish in inflected word form text database indexes. Journal of Information Science 32: 465−479.

Kettunen, K and Airio, E (2006) Is a morphologically complex language really that complex in full-text retrieval? In: Salakoski T et al. (eds.): Advances in Natural Language Processing, LNAI 4139, Springer-Verlag Berlin Heidelberg, 2006, pp. 411–422.

Kettunen, K, Kunttu, T and Järvelin, K (2005) To stem or lemmatize a highly inflectional language in a probabilistic IR environment? Journal of Documentation 61: 476–496.

Kettunen, K, Sadeniemi, M, Lindh-Knuutila, T and Honkela, T (2006) Analysis of EU Languages through Text Compression. In: Salakoski T et al. (Eds.): Advances in Natural Language Processing, LNAI 4139, Springer-Verlag Berlin Heidelberg, 2006, pp. 99–109.

Koskenniemi, K (1996) Finite state morphology and information retrieval. Natural Language Engineering 2: 331− 336.

Kostić, A, Marković, T and Baucal, A (2003) Inflectional Morphology and Word Meaning: Orthogonal or Co-implicative Cognitive Domains. In: Baayen, R H, Schreuder R (eds.): Morphological Structure in Language Processing. Trends in Linguistics, Studies and Monographs 151. Mouton de Gruyter, Berlin, 2003, pp. 1–43.

Koval, S, Beliaeva, L, Kogan, L, Mikhailov, A, Nikolaev, V, Piotrowski, R and Tovmach, Yu (2000) Morphological Representation in PC-Based Text Processing Systems. Literary and Linguistic Computing 15: 131–155.

Kraaij W (2004) Variations on Language Modeling for Information Retrieval. Haag: CTIT Ph. D. series No. 04-62.

Lemur. The Lemur Toolkit for Language Modeling and Information Retrieval. http://www.lemurproject.org/ (visited September 10th, 2006).

Lexin. Svensk-finskt lexicon, http://lexikon.nada.kth.se/sve-fin.shtml (visited in August 2006).

Manning, C D and Schütze, H (1999) Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts.

Mayfield, J and McNamee, P (2003) Single N-gram Stemming. In *Proceedings of Sigir2003, The Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 415−416.

Metzler, D and Croft, W B (2004) Combining the Language Model and Inference Network Approaches to Retrieval. Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval 40: 735–750.

Multitran, http://www.multitran.ru/ (visited in September 2006).

Peters, C (2003) Introduction to the CLEF 2003 Working Notes. http://www.clef-campaign.org/2003/WN_web/00.2%20-%20intro.pdf (visited September 1st, 2005).

Peters, C (2004) What happened in CLEF 2004? Introduction to the working notes. http://clef.isti.cnr.it/2004/working_notes/WorkingNotes2004/CLEF2004WN%20-%20intro.pdf (visited October 20th, 2006).

Petrasi, V, Perelman, N and Gey, F (2003) UC Berkeley at CLEF 2003–Russian Language Experiments and Domain-Specific Cross-Language Retrieval. Working Notes for the CLEF 2003 Workshop 21-22 August, Trondheim, Norway. http://clef.isti.cnr.it/2003/WN_web/29a.pdf (visited October 10th, 2006).

Popovič, M and Willett, P (1992) The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. Journal of the American Society for Information Science 43: 384−390.

Rasmussen, E M (2003) Indexing and Retrieval for the Web. In: Cronin, B (ed.), Annual Review of Information Science and Technology. Volume 37, pp. 91–124.

Russian National Corpus. E-mailed information about noun and adjective distributions, September 1st, 2006. Corpus info available at http://www.ruscorpora.ru/ (in Russian).

Search Engine Showdown. Search Engine Features Chart (Last updated Sep. 17, 2006), http://www.searchengineshowdown.com/features/ (visited October 28th, 2006).

Siegel, S and Castellan, N J Jr. (1988) Nonparametric statistics for the behavioral sciences. Second edition. McGraw-Hill Book Company, New York.

Snowball web page, http://snowball.tartarus.org/ (visited October 20th, 2006).

Sormunen, E 2000. A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Tampere: University of Tampere, Doctoral Thesis. Acta Electronica Universitatis Tamperensis. http://acta.uta.fi/pdf/951-44-4732-8.pdf (visited August 15, 2006).

SWETWOL, http://www2.lingsoft.fi/cgi-bin/swetwol (visited August 2006).

Tiger corpus, http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/ (visited June 7th, 2006).

Tomlinson, S (2004a) Lexical and algorithmic stemming compared for 9 European languages with Humminbird SearchServer™ at CLEF 2003. In Comparative Evaluation of Multilingual Information Access Systems, Springer-Verlag, LNCS #3237, 286−300.

Tomlinson, S (2004b) Finnish, Portuguese and Russian Retrieval with Hummingbird SearchServerTM at CLEF 2004. Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK. http://clef.isti.cnr.it/2004/working_notes/WorkingNotes2004/21.pdf. (visited October 10th, 2006).

Tordai, A and de Rijke, M (2005) Hungarian Monolingual Retrieval at CLEF 2005. Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/tordai05.pdf (visited September 6th, 2006).

Xu, J, and Croft, B (1998) Corpus-based stemming using cooccurrence of word variants. ACM Transactions on Information Systems 16(1), 61−81.

Appendix

Table A1. Word form frequencies of Swedish nouns

FORM	Number of forms	Percentage of forms
INDEF SG NOM	199 846	31.6 %
INDEF PL NOM	86 952	13.7 %
INDEF SG GEN	10 019	1.6 %
INDEF PL GEN	2 392	0.4 %
DEF SG NOM	161 588	25.5 %
DEF PL NOM	64 368	10.2 %
DEF SG GEN	21 940	3.5 %
DEF PL GEN	9 304	1.5 %
SG/PL ambiguity	76 649	12.1 %
SG/PL NOM	73 050	95.3 %
SG/PL GEN	3 599	4.7 %
SUM	**633 058**	**100 %**

Table A2. Case form frequencies for common nouns of German

Case	Number of forms	Percentage of forms
Nominative	54 584	30.5 %
SG.	38 108	
PL.	16 476	
Accusative	47 215	26.4 %
SG.	31 899	
PL.	15 356	
Genitive	21 571	12.1 %
SG.	14 606	
PL.	6 965	
Dative	55 464	31.0 %
SG.	38 916	
PL.	16 548	
SUM	**178 834**	100 %
SG.	123 529	69.1 %
PL.	55 345	30.9 %

Table A3. Case form frequencies for proper nouns of German

Case	Number of forms		Percentage of forms
Nominative	30 048		61.2 %
SG.	29 745		
PL.	303		
Accusative	2 382		4.9 %
SG.	2 293		
PL.	89		
Genitive	3 830		7.8 %
SG.	3 708		
PL.	122		
Dative	12 686		26.1 %
SG.	12 371		
PL.	315		
SUM		**48 946**	**100 %**
SG.		48 117	98.3 %
PL.		829	1.7 %

Table A4. Case form frequencies for Russian nouns

Case	**SG.**	**%**	**PL.**	**%**
Nominative	327637	32.7	76500	25.6
Genitive	236917	23.7	97737	32.7
Dative	53021	5.3	14812	4.9
Accusative	195340	19.5	56929	19.0
Prepositional	89253	8.9	29136	9.7
Instrumental	98789	9.9	24132	8.1
TOTAL	**1000957**	100 %	**299246**	100 %

Table A5. Case form frequencies for Russian adjectives

Case	**SG.**	**%**	**PL.**	**%**
Nominative	76059	31.8	26371	26.8
Genitive	53482	22.4	29989	30.5
Dative	9051	3.8	3983	4.1
Accusative	44079	18.5	17843	18.1
Prepositional	31799	13.3	12347	12.5
Instrumental	24360	10.2	7890	8.0
TOTAL	**238830**	100 %	**98423**	100 %