Is a Morphologically Complex Language Really That Complex in Full-Text Retrieval?

Kimmo Kettunen and Eija Airio

University of Tampere, Department of Information Studies, Kanslerinrinne 1, FIN-33014, University of Tampere, Finland
Kimmo.kettunen@uta.fi, eija.airio@uta.fi

Abstract. In this paper we show that keyword variation of a morphologically complex language, Finnish, can be handled effectively for IR purposes by generating only the textually most frequent forms of the keyword. Theoretically Finnish nouns have about 2,000 different forms, but occurrences of most of the forms are rare. Corpus statistics showed that about 84 – 88 per cent of the occurrences of inflected noun forms are forms of only six cases out of the 14 possible. This number – maximally 2*6 – of keyword's variant forms makes it feasible to try them all in a search. IR results of the frequent keyword form variation coverage were tested with three to twelve keyword variant forms in two test collections, TUTK and CLEF 2003's Finnish material. The results show that the frequent keyword form generation method competes well with the gold standard, lemmatization, with nine and twelve variant keyword forms.

1 Introduction

Various methods for handling of 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 quite rich, cf. [1, 2, 3]. In computational linguistics quarters it seems to have been a common belief that full coverage lemmatization is needed for languages that are morphologically complex [4], even in monolingual single term IR. This belief has been shared also by some IR researchers [5, 6.].

In the same time as simple conflation methods have been used in IR, not much attention has been given to heuristically based language 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. We have earlier shown that our inflectional stem generation method and its further simulated developments compete quite well against the gold standard, FINTWOL, in a best-match IR for Finnish [7, 8]. In this paper we shall further question the need of a full coverage lemmatizer in monolingual IR of a morphologically complex language.

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 411 – 422, 2006.

On a general level, our background motivations can be stated as follows:

1. The average precision and recall (P/R) of retrieval needs to be kept as high as possible without using excessively complex language technology tools; the need of lexicon-based lemmatizers in basic monolingual IR is not as high as often believed even for a morphologically complex language.
2. Performance of new methods introduced 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, cf. e.g., [2, 9, 10, 11]. We have argued in [8] that the performance gained with raw words is quite meager and variable e.g. for Finnish, and thus the increases of performance given for different morphological processing methods are not as positive as they are shown 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. With morphologically complex languages the best retrieval result is usually given by a lemmatizer, such as e.g. TWOL for different languages [4]. This line of argumentation is taken in the present study.

The structure of our paper is following. First we shall show the case distributions of Finnish nouns with corpus statistics. After that the discrepancy between grammatical forms and actual forms in a corpus is pinpointed. After this our research problems are stated, tested and discussed.

2 Case Distribution of Finnish Nouns

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 [13, 14, 15, 16, 17, 18, 19, 20]. Karlsson [18, 19] shows with some semantically distinctive word types, how the case distributions of the words differ in Finnish. A word denoting to a place, *Helsinki*, has besides dominating nominative and genitive singular mainly occurrences of locative cases. A person's name like *Martti* occurs mostly in nominative singular. Same sort of analysis is given e.g. by Kostić et al. [20] 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.

Karlsson [21, pp. 308], citing research of Anneli Pajunen and Ulla Palomäki, presents figures about the distribution of cases for nouns in Finnish. The materials are from four different textual types of the Syntactic archives of Finnish, each comprising 5,000 word form tokens. In this data already the so called grammatical cases, nominative, genitive and partitive, cover 63.5 per cent of the overall distribution of case forms. If some marginal cases are included in this number, the resulting coverage is 67.5 per cent, over two thirds. Out of the other cases eight are so called locative cases (inner locatives: inessive, elative and illative), outer locatives (adessive, allative and ablative) and general locatives (essive and translative), and their share is 30.3 per cent, of which 17.8 per cent are inner locatives. Thus grammatical cases together with the inner locatives make 85.3 per cent of the occurrences of cases in the material.

Räsänen [22] gives a share of 78.2 % for the same six cases based on an analysis of 6 562 word form tokens in three small factual text samples.

When slightly bigger corpora are considered, almost the same distribution is found. In the whole collection of the Syntactic archives of Finnish consisting of 64,391 word form tokens of nouns, 88.2 per cent of the nouns are in the six most frequent cases [23, pp. 1180].

As e.g. Baayen [13, 14] and Biber [15, 16] emphasize, generalizations about linguistic phenomena should be based on large enough corpora. We were able to analyze or get information about two larger Finnish corpora. These were analyzed by a morphological lemmatizer, FINTWOL. We first analyzed the word form types of the inflected index of the TUTK collection [24], 719,011 word form types out of 12,109,779 word form tokens of the database, by running them through the FINTWOL program. All the noun interpretations given by FINTWOL, even ambiguous, were taken into these figures, and thus the figure is an approximation.

Our other and largest analyzed corpus was based on a 32 million word form token HUT corpus, which is one of the largest available corpora for Finnish. Out of the 32 million word form tokens about 11,3 million were analyzed by FINTWOL unambiguously as nouns [25]; these figures from data provided by [26]. In these two different and independent FINTWOL analyses case distributions for the six most frequent cases listed in Table 1 were found.

Table 1. Case distributions in the HUT and TUTK corpora

Cases	Number of noun tokens with one unambiguous TWOL analysis in the HUT corpus	Percentage	Number of noun types with all TWOL analyses in the TUTK corpus	Percentage
Nominative	3,758334	33.14	135,241	26.27
Genitive	2,900884	25.58	109,385	21.24
Partitive	1,428117	12.59	80,158	15.57
Inessive	819,333	7.23	31,007	6.02
Illative	593,513	5.23	41,778	8.11
Elative	520,101	4.59	38,392	7.45
SUM of the six cases	10,020,282/11,339,099	88.36 per cent	435 961/514,795	84. 68 per cent

After stating this, we shall shortly return to the number of forms of Finnish nouns. The usual figure given for the number of grammatical forms of Finnish nouns is about 2,000 [21: pp. 356]. This is achieved by the counting 2 (number) * 13 (cases) * 6 (possessives) * 12 (particles) = 1872. Figure 1,872 is a minimum, maximally the number is slightly over 2,000 if all the variant forms and rare cases etc. are counted for. As we can see from the figures, possessive endings and particles are mostly in charge for the huge number of grammatical word forms, because they can be

combined to every inflected case form. Thus it is of interest to analyze the real occurrence of all the possessives and particles in nouns in a large enough corpus. According to the analyses [26, 29], only 3rd person possessive suffix had a share of 1,83 % when singular and plural are joined, other possessives had negligible distributions. From particles *–kin* had the biggest share, 0,25 %,. other occurrences of particles were negligible.

Relying on these analyses of four different corpora or sub-corpora, it can be argued that about 84 – 88 per cent of the case occurrences of Finnish nouns in (newspaper style) factual text are tokens of six cases only. That is about 43 % of the whole repertoire of Finnish cases. Furthermore particles and possessive endings that make the theoretical number of Finnish nouns so huge are so rare in running texts that they are not of practical importance.

3 Research Problems, Data and Methods

On the basis of these analyses we propose that reasonable or good IR performance can be achieved for Finnish by only taking maximally care of the six cases and their variation in keyword nouns. It is also possible that even with only three cases, nominative, genitive and partitive, quite realistic performance can be expected. The number of cases to be tried out in our tests will thus be three to six. This will mean three to twelve distinct forms of the search keys to be sought for in the database, as singular and plural forms of the cases are distinct. We shall call our method FCG, frequent case (form) generation. Procedures that are tested in this paper are presented and explained in Table 2.

Table 2. Frequent case form procedures to be tested.(* about, accurate number may be bigger in some cases due to variant forms in GEN PL and PTV PL; NOM = nominative, GEN = genitive, PTV = partitive; inner locatives = inessive, elative and illative, PL = plural, SG = singular)

Case forms in the procedure	Number of keyword forms in the procedure	Name of the procedure
NOM-GEN-PTV, only singular	3	FCG_3
NOM-GEN-PTV, singular and plural	6	FCG_6
NOM-GEN-PTV, singular and plural, inner locatives singular only	9*	FCG_9
NOM-GEN-PTV, singular and plural, inner locatives singular and plural	12*	FCG_12

If we contrast the numbers in column two of Table 2 to the theoretical number of Finnish noun forms, we see that in procedure FCG_12 only 0.64 % of the grammatical noun forms are counted for (12/1872). In FCG_3 only 0.16 % of the possible

forms are used (3/1872). Thus the theoretical morphological complexity of Finnish in number of word forms boils down quite a bit and we still believe that reasonable or even good IR performance will be achieved with our procedures.

The emphasis of noun forms in the procedures is due to the well known fact that most of the information content of the texts is carried by nouns, and for that reason mostly nouns are important in queries [27: pp. 169]. Corpus analyses also show that about 35 – 45 per cent of the word tokens in running Finnish texts are nouns. On type level the percentage is 65 – 75 % [28, 29]. Thus the importance of other word classes than nouns in IR is small, and the variation in e.g. verbs does not affect retrieval. Besides nouns we also put adjectives in the FCG procedures in variant case forms; verbs and words of other word classes in topics (besides stop words) were taken into queries in the form they were in the topic.

Our research problem is twofold:

1. Does frequent case form generation of keywords work in IR of a morphologically complex language?
2. If it works, what is the best balance between number of generated keyword form variants and achieved mean average precision in retrieval?

We shall test our case procedures with two collections: TUTK and the Finnish CLEF 2003 material using the InQuery search engine. Both collections have almost the same number of Finnish newspaper articles: TUTK has 53,893 articles from three newspapers from years 1988 – 1992 [24], and CLEF 2003 has 55,344 articles from one newspaper from years 1994 – 1995 [3, 30]. In TUTK [24, 31] we have 30 test topics. The original four relevance levels of the collection are combined in this study: relevance level 3 of TUTK, level of most relevant documents, is called stringent, relevance levels 2 and 3 – level of most relevant and relevant documents – are joined as normal and all the three relevance levels, 1 – 3, are joined as liberal relevance; The rest of the documents, both un-judged and those judged as irrelevant, are taken as irrelevant in this study. In CLEF 2003 we have 60 test topics and binary relevance.

Queries for the test runs were formed partly manually from the topics. After automated initial inflectional stem generation and InQuery query structure generation, the needed case endings were edited to the inflectional stems of the query words, cf. [8]. Thus we simulated carefully the effects of automated rule-based frequent case form generation. Word form generators for Finnish have been implemented since the 1980's [12, 32, 33], but they were not available for this study.

As an example we can take one query from the CLEF 2003 collection. Query #144 for the FCG_3 process is as follows:

#q144 = #sum(#syn(sierra sierran sierraa) #syn(leone leonen leonea) #syn(kapina kapinan kapinaa) #syn(timantti timantin timanttia) #syn(vaikutus vaikutusta vaikutuksen) #syn(kapina kapinan kapinaa) #syn(poliittinen poliittista poliittisen) #syn(epävakaus epävakauden epävakautta) #syn(sierra sierran sierraa) #syn(leone leonen leonea) #syn(timanttiteollisuus timanttiteollisuuden timanttiteollisuutta));

The queries are of the form #SUM(#SYN() #SYN()…), and thus they are strongly structured [34]. Morphological variant forms of the keyword are treated as synonyms of the key, and InQuery treats them as instances of one key [35, 36].

Results of the FCG procedures are compared to results of FINTWOL lemmatization and Snowball stemming, which have earlier been shown to work well in Finnish best-match retrieval [3, 7, 8]. Results with plain keywords are shown for comparison as a worst case performance.

4 Results

Results with the TUTK collection are presented in Table 3. Differences in the tables are actual percentages, not relative. (* In the Plain method keywords are taken as such straight from the topics in the forms they happen to be there.)

Table 3. Results of test runs in the TUTK collection on three relevance levels

Method	Liberal relevance Mean average precision (per cent) - interpolated	Normal relevance Mean average precision (per cent) - interpolated	Stringent relevance Mean average precision (per cent) - interpolated
FINTWOL – lemmatized index, compounds split in the index	37.8	35.0	24.1
FCG_12, inflectional index	32.7 (-5.1)	30.0 (-5.0)	21.4 (-2.7)
FCG _9, inflectional index	32.4 (-5.4)	29.6 (-5.4)	21.3 (-2.8)
FCG _6, inflectional index	30.9 (-6.9)	28.0 (-7.0)	21.0 (-3.1)
Snowball, stemmed index	29.8 (-8.0)	27.7 (-7.3)	20.0 (-4.1)
FCG _3, inflectional index	26.4 (-11.4)	23.9 (-11.1)	18.9 (-5.2)
*Plain, inflectional index	19.6 (-18.2)	18.9 (-16.1)	12.4 (-11.7)

Results from the CLEF 2003 runs are in Table 4. Non-interpolated figures for FINTWOL, Snowball and Plain are from [3] and they are shown for comparison.

Results are quite similar in both collections, although FCGs perform better in the CLEF collection overall. FCG_3 performs very poorly in CLEF 2003 collection, and it outperforms plain words only slightly, while in TUTK the difference between FCG_3 and Plain keywords is clear. FCG_6 performs much better in CLEF 2003 than

Table 4. Results of test runs in CLEF 2003 collection

Method	Mean average precision (per cent) - interpolated	Mean average precision (per cent) - non-interpolated
FINTWOL - lemmatized index, compounds split in the index	37.6	50.5
- lemmatized index, compounds not split in the index	34.7 (-2.9)	47.0 (-3.5)
Snowball, stemmed index	35.8 (-1.8)	48.5 (-2.0)
FCG_12, inflectional index	34.0 (-3.6)	46.4 (-4.1)
FCG_9, inflectional index	33.7 (-3.9)	46.1 (-4.4)
FCG_6, inflectional index	30.1 (-7.5)	41.5 (-9.0)
FCG_3, inflectional index	24.2 (-13.4)	32.6 (-17.9)
Plain, inflectional index	22.7 (-14.9)	31.0 (-19.5)

FCG_3, but still hangs 6.5 – 9 per cent-units below Snowball and FINTWOL. In neither collection FCG_12 brings much gain to FCG_9, it is only 0.1 – 0.4 per cent-units better than FCG_9.

Our best case procedures, FCG_9 and FCG_12 perform well in both collections and they are only 3.6 – 4.4. per cent behind FINTWOL in CLEF 2003 and 2.8 - 5.4 per cent in TUTK depending on the relevance level.

We tested the statistical significance of the differences between the best methods in both collections using the Friedman test. Tested methods were FINTWOL, Snowball, FCG_12, FCG_9 and FCG_6. Although FCGs do not outperform Snowball in CLEF 2003 on any level, the differences between FCG__9, FCG__12, FCG__6 and Snowball are not statistically significant. The difference between FINTWOL using split compound index and FCG__6 was statistically significant ($p = 0.005$) in CLEF 2003. Difference between FINTWOL with compounds not split and FCG__6 was also statistically significant ($p = 0.02$). Differences between FINTWOL, FCG__9 and FCG__12 were not statistically significant in CLEF 2003.

In the TUTK collection the differences were more often statistically significant. Table 5 presents the statistical differences in the TUTK collection when the Friedman test was used. Only significant differences are listed.

Table 5. Statistically significant differences between the best methods in the TUTK collection

	Liberal relevance	Normal relevance	Stringent relevance
FINTWOL	> ALL	> ALL	> ALL but FCG_9
FCG_12	---	> Snowball p = 0.02	---
FCG_9	---	> Snowball p = 0.01	---
FCG_9	---	> FCG_6 p = 0.03	---

5 Discussion

According to the results it seems that the nine forms of FCG__9 are optimal for the search in both collections. By adding three more keyword forms to the query, only marginal gains are achieved. This is shown more clearly in Figure 1, where mean average precisions from both collections are compared to the number of variant keyword forms (for TUTK only the liberal relevance level curves are shown).

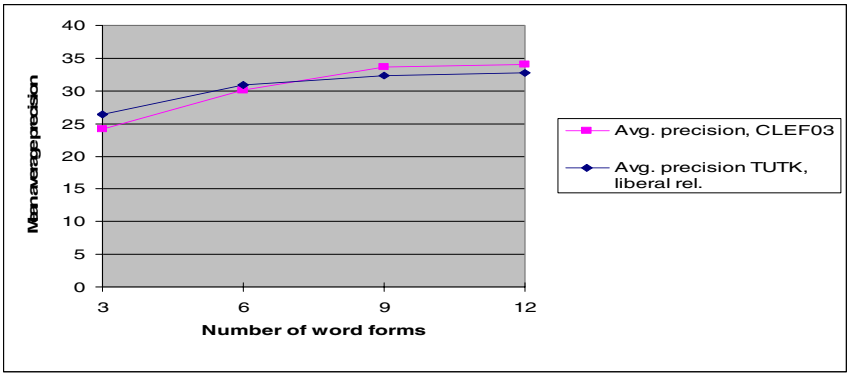

Fig. 1. Number of variant keyword forms and mean average precisions of FCG procedures

The figure shows that the mean average precisions of the queries almost stop rising in both collections after nine forms. This may be due to two reasons: either there is no large gain to be achieved with any of the added forms after nine forms, or the three additional forms in process FCG__12 are not the right ones (plural forms of inner locatives). From the results of frequency analysis it is possible that also singular forms of two outer locative cases (adessive and allative) or general locative cases (essive and translative) could be better forms to be used in addition to the nine forms. This was not tested any further. It is possible that a slight improvement of average precision over FCG__12 can be achieved by using different case forms beyond the

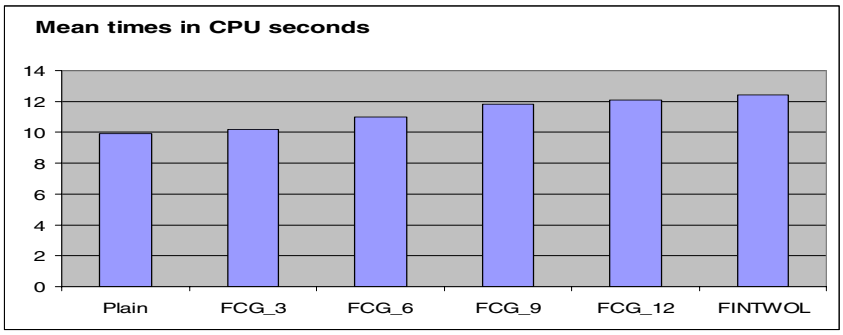

Fig. 2. Mean runtimes in CPU seconds with 60 queries of CLEF 2003

nine forms of the procedure FCG__9. In single queries this is at least certainly true, as the semantic types of the keywords may favor some other cases than those used in our FCGs.

Our experiments showed also that the use of three to twelve full form variants of each keyword is computationally tractable. Results of the CPU time tests for the whole set of 60 queries in CLEF 2003 are shown in Figure 2 (mean CPU seconds of five consecutive runs, system time + user time of Unix's *time* function added together; test system was a Sun Sparc Station with two 1,015 GHz processors and 4 GB memory under a timesharing load of a few concurrent users).

As can be seen from Figure 2, adding keyword forms to FCG procedures does not increase runtimes very much. The increase in mean CPU time is only about 20 % when maximal 12 keyword forms are used instead of three. Plain unprocessed keywords are fastest to run, and slightly surprisingly FINTWOL queries are not faster to run than FCG processes.

In earlier publications we compared other types of morphological variant handling to lemmatization. Inflectional stem generation was found to be almost as effective in average precision as lemmatization, but it resulted in slowly processed very large queries [7]. When inflectional stems were enhanced with regular expressions which restricted the choice of possibly matching words from the index, queries were faster to run and more manageable in size, but they had lower average precision [8]. When compared to these results, FCG style of keyword handling seems to be the most optimal in both average precision and runtime for inflectional indexes. The best FCG processes achieve about 86 % of the best gold standard results in TUTK and about 90 % of the best results in CLEF 2003. If best FCGs are compared to FINTWOL in CLEF 2003 with non-split compound index, FCG_12 achieves about 98 % of the mean average precision of FINTWOL and FCG_9 about 97 %. As this performance level is achieved without runtime penalties in inflected indexes, the results can be considered very good.

Thus the use of frequent case form generator as shown in this paper would be a viable alternative to be used in real query systems, such as web search engines, which do not many times have any means for handling the morphological variation of keywords in many languages. It should also be noted that the mean number of keywords given by a web-user is less than three [37]. We had long queries made out of long

topics, but they ran fine. Our method has also other advantages besides a reasonable average P/R: it works with inflected form indexes and will not suffer as much from out of vocabulary words as lemmatizers; FCGs should also be simple to implement for new languages, even if they are language specific and need linguistic expertise.

6 Conclusion

The purpose of this paper was to evaluate use of only the most frequent keyword forms in a monolingual full-text retrieval of a highly inflectional language, Finnish. The forms to be used in retrieval were first analyzed from several text corpora of variable sizes. Corpus analysis showed that six cases 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. This shows that, while a language may in principle be morphologically complex, in practice it is much less so. Based on this finding, four different simulated frequent case form generation procedures (FCGs) were tested in two different full-text collections, TUTK and CLEF 2003.

The results show that frequent case form generation works in full-text retrieval 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, achieved about 86 % of the best average precisions of FINTWOL in TUTK and about 90 % in CLEF 2003. The runtimes of the FCG queries were also shown to be comparable to those of the other methods. Thus the 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.

It was also shown that corpus statistics of inflectional form distributions were useful for choosing a limited set of basic case forms to cover in a language technology application of a single highly inflectional language. This finding together with general knowledge about token frequency distributions suggests that the method is suitable for other languages too, and thus our results need not be language specific only. Morphologically less complex languages may be served with simpler FCGs with quite few word forms. As the presented method is easily testable for any language of even modest morphological complexity, it can be evaluated on a language by language basis.

Acknowledgements

We wish to thank Mr. Mathias Creutz for providing the data of the HUT corpus. Kal Järvelin has commented earlier drafts of this paper insightfully. 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 Snowball Finnish stemming algorithm is available at http://snowball.tartarus.org/algorithms/finnish/stemmer.html. (Accessed November 28th, 2003).

References

1. Popovič, M., Willett, P.: The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. Journal of the American Society for Information Science 43 (1992) 384–390
2. Hollink, V., Kamps, J., Monz, C.,de Rijke, M.: Monolingual Document Retrieval for European Languages. Information Retrieval 7 (2004) 33–52
3. Airio, E.: Word Normalization and Decompounding in Mono- and Bilingual IR. Information Retrieval (2005), to appear**.**
4. Koskenniemi, K.: Finite State Morphology and Information Retrieval. Natural Language Engineering 2 (1996) 331 – 336
5. Galvez, C., Moya-Anegón, F., Solana, V. H.: Term Conflation Methods in Information Retrieval. Non-linguistic and Linguistic Approaches. Journal of Documentation 61 (2005) 520 – 547.
6. Jacquemin, C., Tzoukerman, E.: 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 (1999) 25 – 74
7. Kettunen, K.: Developing an Automatic Linguistic Truncation Operator for Best-match Retrieval in Inflected Word Form Text Database Indexes. To appear in Journal of Information Science 32 (2006)
8. Kettunen, K., Kunttu, T., Järvelin, K.: To Stem or Lemmatize a Highly Inflectional Language in a Probabilistic IR Environment? Journal of Documentation 61 (2005) 476 – 496
9. Braschler, M., Ripplinger, B: How Effective is Stemming and Decompounding for German Text Retrieval? Information Retrieval 7 (2004) 291 – 316
10. Mayfield, J., McNamee, P. Single N-gram Stemming. In: Proceedings of Sigir2003, The Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2003) 415–416
11. Tomlinson, S.: Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer™ at CLEF 2003. Availabe at: http://clef.iei.pi.cnr.it/2003/WN_web/19.pdf (accessed April 28th, 2004)
12. Koskenniemi, K: A System for Generating Finnish Inflected Word Forms. In Karlsson, F. (ed.): Computational Morphosyntax. Report on research 1981 – 84. Publications of the Department of General linguistics, University of Helsinki. No. 13 (1985) 63 – 80
13. Baayen, R. H.: Statistical Models for Word Frequency Distribution. Computers and the Humanities 26 (1993) 347 – 363
14. Baayen, R. H.: Word Frequency Distributions. Kluwer Academic Publishers, Dordrecht Boston London (2001)
15. Biber, D.: Representativeness in Corpus Design. Literary and Linguistic Computing 8 (1993) 243 – 257
16. Biber, D.: Using Register-diversified Corpora for General Language Studies. Computational Linguistics 19 (1993) 219 – 241
17. Manning, C. D., Schütze, H.:. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts (1999)
18. Karlsson, F.: Frequency Considerations in Morphology. Zeitsschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 39 (1986) 19 – 28
19. Karlsson, F.: Defectivity. In: Booij G. et al. (eds.): Morphology. An International Handbook on Inflection and Word-Formation. Volume 1. Walter de Gruyter, Berlin (2000) 647 – 654

20. Kostić, A., Marković, T., Baucal, A:. 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) 1 – 43
21. Karlsson, F.: Suomen kielen äänne- ja muotorakenne. WSOY, Helsinki (1983)
22. Räsänen, S.: Havaintoja suomen sijojen frekvensseistä. (Observations of frequencies of the Finnish cases) Sananjalka 21 (1979) 17–43
23. Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T.R., Alho, I. Iso suomen kielioppi.: Suomalaisen Kirjallisuuden Seura, Helsinki (2004)
24. Sormunen, E.: A Method for Measuring Wide Range Performance of Boolean Queries in Full-text Databases. Acta Universitatis Tamperensis 748, Tampere (2000)
25. Creutz, M., Linden, K.: Morpheme Segmentation Gold Standards for Finnish and English. Helsinki University of Technology. Publications in Computer and Information Science. Report A77. Espoo (2004)
26. Creutz, M.: Two E-mails, May 17, 2005
27. Baeza-Yates, R., Ribeiro-Neto B.: Modern Information Retrieval. Addison Wesley, USA 1999
28. Saukkonen, P., Haipus, M., Niemikorpi A., Sulkala, H:. Suomen kielen taajuussanasto. (A Frequency Dictionary of Finnish). WSOY, Helsinki (1979)
29. Kettunen, K.: Sijamuodot haussa - tarvitseeko kaikkea hakutermien morfologista vaihtelua kattaa? Ms. Sci. Thesis, University of Tampere, Department of Information Studies (2005)
30. Peters, C.: Introduction to the CLEF 2003 Working Notes. Available at http://www.clef-campaign.org/2003/WN_web/00.2%20-%20intro.pdf. Accessed 1st September 2005.
31. Sormunen E.: The Effectiveness of Free-text Searching in Full-text Databases Containing Newspaper Articles and Abstracts. Research Publications 790. Technical Research Centre of Finland., Espoo (In Finnish, English abstract) (1994)
32. Holman, E.: Finnmorf: A Computerized Research Tool for Students of Finnish Morphology. Computers and the Humanities 22 (1988) 165 – 172
33. Lassila, E.: Suomen kielen sanamuodot taivuttava ohjelma FORMO. In: Mäkelä, M. Linnainmaa , S., Ukkonen, E (eds.): STeP-88. Invited Papers. Contributed Papers: Applications. Helsinki: Finnish Artificial Intelligence Society (1988) 118 – 126
34. Kekäläinen, J.: The Effects of Query Complexity, Expansion and Structure on Retrieval Performance in Probabilistic Text Retrieval. Acta Universitatis Tamperensis 678 , Tampere(1999)
35. Allan, J., Callan, J., Croft, B., Ballesteros, L., Byrd, D., Swan, R., Xu, J.: INQUERY Does Battle with TREC-6. In: Voorhees, E. & Harman, D. (Eds.) Proceedings of the TREC 6 Conference. (1997). Available from http://trec.nist.gov/pubs/trec6/t6_proceedings.html. Accessed Nov 15, 2005
36. Broglio, J., Callan, J., Croft, W.B.: INQUERY System Overview. In: Proceedings of the TIPSTER text program (Phase I). San Francisco, CA: Morgan Kaufmann Publishers. (1994)
37. Jansen, B., Spink, A., Sarasevic, T.: Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. Information Processing & Management 36 (2000) 207–227