

# Covering the morphological variation of Finnish query nouns in a probabilistic best-match IR environment

**Kimmo Kettunen**

Department of Information Studies, University of Tampere  
Tampere, Finland  
Kimmo.kettunen@uta.fi

## Abstract

Effects of three different morphological methods - normalisation, stemming and stem production - for Finnish are compared in a probabilistic best-match IR environment (InQuery). Evaluation is done using two different relevance scales: a four-point relevance scale and a binary relevance scale. Results show that stem production, a lighter method than morphological normalisation, compares well with normalisation in a best-match IR environment. Differences in performance between stem production and normalisation are small and they are not statistically significant. It is also shown that hitherto a rather neglected method of morphological processing for Finnish, stemming, works better than thought although the used stemmer – a Porter stemmer implementation – is far from optimal for a morphologically complex language like Finnish.

## 1. Introduction

Due to the rich morphology of Finnish inflected word forms or ad hoc truncated general forms are not considered as good search keys for information retrieval in Finnish documents. Some kind of morphological processing of the search terms is needed for getting satisfying results. Four different language dependent methods exist for handling of morphological variation: normalisation, stemming, stem production and full form generation of words. Also a variety of language independent methods, such as n-grams and other fuzzy matching methods exist. Some of the language dependent methods can be considered better for a certain morphological type of language. Some do not fit a specific language at all. In the case of Finnish, e.g. stemming has not been considered a suitable method and full form generation is clearly unpractical due to the number of different possible surface forms. The computational burden of 28 (2\*14) case forms in singular and plural would already be high for the retrieval system.

Automatic stem production and normalisation programs for Finnish have been implemented since early 1980's and they have been used in information retrieval programs. The most

prominent stem programs have been Finstems (Koskenniemi, 1985) and Hahmotin by Kielikone Ltd. (Alkula, 2000). There are also at least three different morphological normalisers for Finnish: Fintwol (Koskenniemi, 1983), Morfo (Jäppinen and Ylilammi, 1986) and Ment (Blåberg, 1994). General overall evaluation of the value of normalisation and stem production programs for IR in a Boolean retrieval environment has been published recently in (Alkula, 2000, 2001). Tested programs were Finstems, Hahmotin, Fintwol and Morfo. Morphological normalisation and stem production were tested with different types of indexes (normalised index, with compounds split and not split, and inflected index) and also manual truncation of the search terms by a seasoned user was tried out. Two of these methods, normalisation and stem generation, are tried out in this study in a best-match system. Two other methods, namely stemmed query words and plain unprocessed query words as such are also tested.

What the used morphological methods should be in a probabilistic best-match IR environment, such as InQuery, has been an open question. Kunttu (2003) is the first one to test these methods for a best-match probabilistic retrieval system for Finnish. Mayfield and McNamee (2003) have tried stemming and n-gram methods for a variety of languages, which also include

Finnish. In this paper, we show performance results of normalisation, stem generation and stemming for Finnish.

The research problems of this paper are as follows:

- How do normalisation and stem production compare in a probabilistic best-match environment?
- Is stemmer a realistic alternative for handling of Finnish morphology for IR?

In Section 2, we present data and methods. Section 3 presents results; Sections 4 and 5 contain discussion and conclusions.

## 2. Data and methods

The tests of this study were conducted in the Information Retrieval Laboratory of the Department of Information Studies, University of Tampere. Actual searches were conducted with a probabilistic partial match system, InQuery, version 3.1 (Broglio & al, 1996).

The test collection, TUTK, contains a full-text database of newspaper articles published in three Finnish newspapers in 1988 – 1992. The newspapers are Aamulehti (a general newspaper), Keskiuomalainen (a general newspaper) and Kauppalehti (an economics oriented five day newspaper), and the database consists of 53 893 articles. Articles represent different sections of the newspapers, mostly economics (from all sections of Kauppalehti, some 16 000 articles), and foreign and international affairs (Aamulehti, some 25 000 articles) and articles from all sections of Keskiuomalainen (some 13 000 articles). (Sormunen, 2000).

The index of the database contains all the word form types of the texts. No stop word lists are used to exclude any words from the index. Articles of the database are fairly short, about 202 words on average. Typical text paragraphs are two or three sentences in length. The set of topics consists of 30 topics (Sormunen, 2000). Topics are long: the mean length of the original topics is 17.4 words. When stop words are omitted, mean length is 15.06 words per topic.

### 2.1. Relevance levels of TUTK

Sormunen (2000, 63) describes the relevance levels of TUTK test collection. A four-point scale is used and its interpretation is in table 1.

**Table 1.** Relevance levels of TUTK interpreted.

Relevance level	Interpretation
0	Totally off target
1	Marginally relevant, refers to the topic but does not convey more information than the topic description itself
2	Relevant, contains some new facts about the topic
3	Highly relevant, contains valuable information, the article's main focus is on the topic.

In this study, queries were both performed on separate relevance levels from 1 - 3 and binary relevance level. Binary relevance level was created from the original four levels by combining levels 2 and 3 to Polar1 (relevant); levels 0 and 1 were considered irrelevant. Separate relevance levels are referred to in performance tables as Rel1, Rel2 and Rel3, and binary level as Polar1.

### 2.2. Methods

The main purpose of this study was to test three different morphological methods in a best-match IR environment. Two of the programs, Fintwol<sup>1</sup> and Snowball<sup>2</sup>, were gotten from external sources. They represent normalization and stemming approaches in this study.

In early 1990's we implemented a noun stem

<sup>1</sup> Fintwol is an implementation of the two-level model for Finnish by Lingsoft (<http://www.lingsoft.fi>), its original contribution is (Koskenniemi, 1983).

<sup>2</sup> Snowball (Porter, 2001) is a language for defining stemmers. A stemmer for Finnish has been produced according to its ideas and based on linguistic description of Finnish. Algorithm of the stemmer is described on web page The Finnish stemming algorithm.

production program for Finnish which was named Stemma. It is described in more detail in Kettunen (1991a, 1991b). Later another version was implemented. The original program is called now MaxStemma and the later version MinStemma. The names imply that programs produce noun stems differently: MaxStemma maximizes the number and length of stems; MinStemma minimizes the number of stems and produces shorter stems. Stem production of both programs is linguistically motivated, and no ad hoc stems for IR purposes are generated. In this study MaxStemma was used for stem generation.

Before going on to the test procedure, a short discussion of terms **normalisation**, **stem production** and **stemming** is necessary.

In linguistics stem is defined as a basic unit from which inflected word forms are generated by adding affixes (Matthews, 1991). This is also the starting point of the approach taken in the development of MaxStemma: stems are generated from input basic forms for further use. Stemming, then, as it is used in the IR literature has different goals: a stemmer analyses inflected word forms and produces conflated stems that can be used in information retrieval. While linguistic stems are linguistically motivated basic elements of surface form production, stems in IR may be almost anything: stems or roots or ad hoc truncated forms that can be adjusted to varying IR needs.

**Lemmatisation** or **normalisation** is usually used to describe the process when inflected word form and its (dictionary) basic form are related to each other with an algorithm. Stemming can be seen as a simpler variant of lemmatisation (e.g. Jacquemin and Tzoukerman, 1999): some stemming programs may be very sophisticated and use full dictionaries (e.g. Krovetz 2000) while others (or most) are simpler and do not usually use dictionaries. (Hull, 1996, Harman 1991, Baeza-Yates and Ribeiro-Neto, 1999, Frakes 1992, Koskenniemi, 1983, 1985, Paice, 1996).

Thus **stem production** will be used in this study to mean production of inflectional stem variants. **Stemming** is used for Porter and Lovins type of approaches, and **normalisation** for Fintwol type of approaches. The demarcation line between stemming and normalisation, however, is not clear cut. In e.g. Kraaij & Pohlman (1996) and Braschler & Ripplinger (2003) several different types of stemmers are introduced for Dutch and

German, and it would be fair to characterize some of them as normalisation programs, while their analysis is based on large dictionaries and is also linguistically motivated. Such a program is also stemmer of Krovetz (2000), while it uses a full dictionary (The Longman Dictionary of English) to check the proposed stemming before accepting it.

### 2.3. The query process

For the testing of the different methods, an automatic IR laboratory procedure was built. The basic query process consists of the following steps when stem production is tested:

1. Normalisation of the query words (from topic descriptions) with Fintwol.
2. Basic InQuery query generation resulting in a slightly structured query with InQuery's #sum and #syn operators.
3. Elimination of stop words in the query according to a stop list of Finnish.
4. Generation of stems for final query words with MaxStemma.
5. Grepping of the database index with stems which will conclude in final query formulation; grepping simulates term truncation which is not supported in InQuery.
6. InQuery retrieval and its evaluation.

When normalisation is used, only steps 1, 2, 3 and 6 are in use. When stemming is used, step 1 consists of stemming of the query words. After that steps 2, 3 and 6 are used. When plain topic words are used as query words, only steps 2, 3 and 6 are used.

In table two differences of database indexes for different used methods are described briefly.

**Table 2.** Types of indexes with different methods.

Used method for query words	Type of index
plain word	inflected index (no processing)
stemmed word	stemmed index (stemmed with Snowball)
Stems	inflected index (no processing)

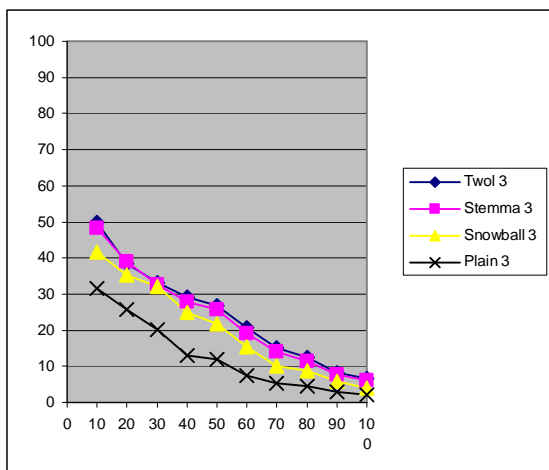
normalised word form	normalised index (normalised with Fintwol, compounds not split)
----------------------	--

### 3. Results

In this section, we show the main results of this study. In tables 3 and 4 average precisions of different methods are presented on relevance level 3 and binary relevance level Polar1. In figures 1 and 2 eleven point P-R curves of the same relevance levels are shown.

**Table 3.** Average precision results on relevance level 3, differences between methods in percentage units.

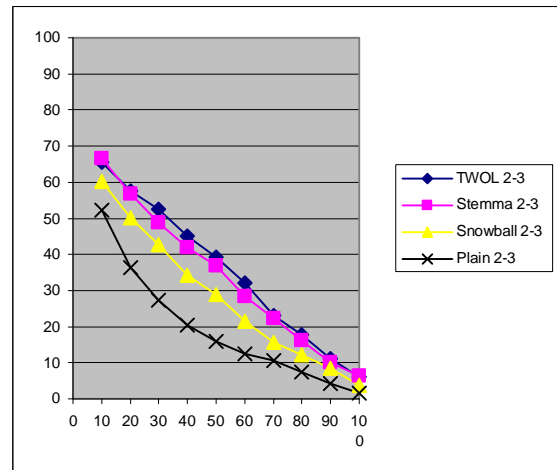
Rel 3	Fintwol	MaxStem ma	Snowball	Plain
Avg. precision	24.1	23.2	20.0	12.4
Fintwol		-0.9	-4.1	-11.7
MaxStem ma			-3.2	-10.8
Snowball				-7.6



**Figure 1.** P-R curves on 11 recall levels on relevance level 3.

**Table 4.** Average precision results on binary relevance level, differences between methods in percentage units.

Polar1	Fintwol	MaxStem ma	Snowball	Plain
Avg. precision	35.0	33.5	27.7	18.9
Fintwol		-1.5	-7.3	-16.1
MaxStem ma			-5.8	-15.6
Snowball				-8.8



**Figure 2.** P-R curves on 11 recall levels on binary relevance level.

The rest of the results are shown briefly in table 5.

**Table 5.** Average precision results on relevance levels 2 and 1, differences between methods in percentage units.

Rel 2	Fintwol	MaxStemma	Snowball	Plain
Avg. precision	21.0	20.1	17.5	12.2
Fintwol		-0.9	-3.5	-8.8
MaxStemma			-2.6	-7.9
Snowball				-5.3
Rel 1	Fintwol	MaxStemma	Snowball	Plain
Avg. precision	11.4	11.9	8.7	5.5
Fintwol		0.5	-2.7	-5.9
MaxStemma			-3.2	-6.4
Snowball				-3.2

### 3.1 Statistical testing of the differences

The statistical testing of the differences between methods was done using the Friedman test (Siegel and Castellan, 1988, modifications by Conover, 1980). All used methods were evaluated against each other with different relevance levels. Significant differences at levels 0.001 and 0.01 are shown in table 6.

**Table 6.** Differences between methods significant at levels 0.001 and 0.01 using the Friedman test. Differences at 0.001 level shown in bold.

Comparison	Significant difference
Fintwol vs. MaxStemma	None on any relevance level
Fintwol vs. Snowball	<b>Rel3, Rel2, Rel1, Polar1</b>
Fintwol vs. plain words	<b>Rel3, Rel2, Rel1, Polar1</b>
MaxStemma vs. Snowball	<b>Rel2, Rel1, Polar1</b>
MaxStemma vs. plain words	<b>Rel3, Rel2, Rel1, Polar1</b>
Snowball vs. plain words	<b>Rel3, Rel1, Polar1</b>

## 4. Discussion

Two research problems were formulated for this study. Firstly, we were interested to see how normalisation and stem production compare in a probabilistic best-match environment. Second question was whether a stemmer is a realistic alternative for the handling of Finnish morphology for IR.

For the first question it was shown, that normalisation performed maximally 1.5 percentage units better than stem production on average. However, the differences between Fintwol and MaxStemma were not statistically significant on any relevance level

Sparck Jones (1974) introduced measures for practical importance of statistical differences between methods. If differences between two methods are statistically significant, their practical differences can be evaluated as a rule of thumb followingly:

- If the difference between methods is less than 5 %, the practical difference is not noticeable
- If the difference between methods is 5 – 10 %, the practical difference is noticeable.
- If the difference between methods is >10 %, the practical difference is material

Differences between Fintwol and MaxStemma were not statistically significant and their practical

differences are not noticeable on any relevance level. Average precisions on separate relevance levels 1 – 3 were not noticeable between Fintwol and Snowball and MaxStemma and Snowball. On binary relevance level Polar1 differences between Fintwol and Snowball and MaxStemma and Snowball were noticeable (7.3 % and 5.8 %, respectively).

Differences between Fintwol and plain words and MaxStemma and plain words were material on relevance level 3. On the same relevance level difference between the stemmer and plain words was noticeable.

On relevance level 2 differences between all the morphological methods against plain words was noticeable. On relevance level 1 differences between Fintwol and plain words and MaxStemma and plain words were noticeable. On binary relevance level Polar1, differences between Fintwol and plain words and MaxStemma and plain words were material. Difference between Snowball and plain words on this relevance level was noticeable.

Is stemmer, then, a realistic alternative to handling Finnish morphology for IR? Mayfield and McNamee (2003) report their experiments on CLEF 2002 collection with eight different languages. Languages include also Finnish and their tested methods include Snowball stemmer, plain words and different types of n-grams. They get their best results for Finnish with 4-grams. Snowball works worse than 4-grams, but it is the second best method for Finnish and the stemmer gets over 10 % units better average results than in our study. Airio et al (2003) studied the use of Snowball stemmer as an indexing method in their CLEF experiment. Their conclusion was that indexing with Snowball was not a good method for CLIR purposes. Results with the stemmed indexes of Finnish material were about 15 % units lower than results with morphological normalisation.

Earlier stemming has not been really used for Finnish IR. Considering Snowball's origin and stemming style (suffix stripping, which is not very suitable for a heavily inflected language, cf. Frakes, 1992, Pirkola, 2001) its performance is better than thought although clearly below performance of normalisation and stem production. It is possible, that a well done linguistically motivated stemmer could be enough for handling the morphology of Finnish for IR

purposes. In e.g. Kraaij and Pohlman (1996) and Braschler and Ripplinger (2003) several different types of stemmers are introduced for Dutch and German and some of them are fairly sophisticated, while their analysis is based on large dictionaries and is also linguistically motivated. Such a program is also stemmer of Krovetz (2000). It is thus assumable that a more sophisticated stemmer for Finnish would also perform better. But as mentioned in section 2.2, the demarcation line between a stemmer and a normaliser would be hard to draw if the stemmer is using full dictionaries.

## 5. Conclusion

We have tested three different methods for handling the morphological variation of Finnish query words in a probabilistic best-match environment. Also, queries with totally unprocessed query words were tried out.

Main results of this study were the following:

- Differences between stem production and normalisation were small and not statistically significant in the tested environment; their practical differences were not noticeable on any relevance level. Only in individual queries there were greater differences, but these differences were not analysed thoroughly.
- The stemmer for Finnish performed better than thought, although the stemmer was not optimal. Still the difference between the stemmer and normalisation or stem production was mostly statistically and practically significant.

## 6. Bibliographical References

Eija Airio, Heikki Keskustalo, Turid Hedlund and Ari Pirkola 2003. Multilingual Experiments of UTA at CLEF 2003. The impact of different merging strategies and word normalizing tools. In *Results of the CLEF 2003 Evaluation Campaign*, Carol Peters, Francesca Borri, eds., Cross-Language Evaluation Forum, Italy, pp. 13 – 18.

Riitta Alkula 2000. *Merkkijonoista suomen*

kielen

sanoiksi. Acta Universitatis Tamperensis 763. Available at <http://acta.uta.fi/pdf/951-44-4886-3.pdf>. Visited March 31<sup>st</sup> 2004.

Riitta Alkula 2001. From Plain Character Strings to Meaningful Words: Producing Better Full Text Databases for Inflectional and Compounding Languages with Morphological Analysis Software. *Information Retrieval* 4, pp. 195 – 208.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto 1999. *Modern information retrieval*. ACM Press, New York.

Olli Blåberg, 1994. *The Ment Model – Complex States in Finite State Morphology*. Reports from Uppsala University, Dept. of Linguistics. RUUL 27.

Martin Braschler and Bärbel Ripplinger 2003. Stemming and Decomposing for German Text Retrieval. In *Advances in Information Retrieval*. 25<sup>th</sup> European Conference on IR Research, ECIR 2003, Pisa, Italy, pp. 177 – 192.

John Broglio, James P. Callan, Bruce Croft and Daniel W. Nachbar 1995. Document Retrieval and Routing Using the INQUERY System. In *Proceedings of the Third Text Retrieval Conference (TREC-3)*. Gaithersburg, MD: National Institute of Standards and Technology, special publication 500-225, pp. 29 – 38.

W. J. Conover, 1980. *Practical Nonparametric Statistics*. 2<sup>nd</sup> edition. John Wiley & Sons, New York.

William B. Frakes 1992. Stemming Algorithms. In W. B. Frakes, & R. Baeza-Yates, eds., *Information Retrieval. Data Structures and Algorithms*, pp. 131 – 160.

William B. Frakes and Ricardo Baeza-Yates (eds.) 1992. *Information Retrieval. Data Structures and Algorithms*. Prentice Hall.

Donna Harman 1991. How effective is Suffixing? *Journal of the American Society for*

*Information Science* 42 (1), pp. 7 - 15.

David A. Hull 1996. Stemming Algorithms: a Case Study for Detailed Evaluation. *Journal of the American Society for Information Science* 47 (1), pp. 70 – 84.

Christian Jacquemin and Evelyne Tzoukerman 1999. NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In *Natural Language Information Retrieval*, T. Strzalkowski (ed.), Kluwer Academic Publishers, pp. 25 – 74.

Harri Jäppinen and Matti Ylilampi 1986. Associative Model of Morphological Analysis: an Empirical Inquiry. *Computational Linguistics* 12(4), pp. 257 – 272.

Fred Karlsson 1985. *Computational morphosyntax*. Report on research 1981 - 84. Publications of the Department of General linguistics, University of Helsinki. No. 13.

Kimmo Kettunen 1991a. Doing the Stem Generation with Stemma. In *Papers from the Eighteenth Finnish Conference of Linguistics*, J. Niemi (ed.) Kielitieteellisiä tutkimuksia, Joensuu yliopisto, N:o 24, 1991, pp. 80 – 97

Kimmo Kettunen, Kimmo, Stemma 1991b. A Robust Noun Stem Generator for Finnish. *Humanistiske Data* 1/91, pp. 26 – 31.

Kimmo Koskenniemi 1985. FINSTEMS: a Module for Information Retrieval. In Karlsson 1985, pp. 81 – 92.

Kimmo Koskenniemi 1983. *Two-Level Morphology: a General Computational Model for Word-form Recognition and Production*. Publications of the Department of General linguistics, University of Helsinki. No. 11.

Kraaij, Wessel - Renée Pohlmann 1996. Viewing stemming as recall enhancement. In *Proceedings of the 19<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zürich, Switzerland, 40 – 48.

Krovetz, Robert 2000. Viewing morphology as an

- inference process. *Artificial intelligence* 118, pp. 277 – 294.
- Kunttu, Tuomas 2003. Perus- ja taivutusmuotohakemiston tuloksellisuus todennäköisyyksiin perustuvassa tiedonhaku-järjestelmässä. Informaatiotutkimuksen pro gradu -tutkielma. Informaatiotutkimuksen laitos, Tampereen yliopisto. (M.Sc. Thesis).
- Matthews, P. H. 1991. Morphology. Second edition. Cambridge University Press.
- Mayfield, James and Paul McNamee 2003. Single N-gram Stemming. In *Proceedings of Sigir2003*, The Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 415 – 416.
- Pirkola, Ari 2001. Morphological Typology of Languages for IR. *Journal of Documentation*, 57: 330 – 348.
- Martin F. Porter 2001. Snowball: A language for stemming algorithms. Available at <http://snowball.tartarus.org/texts/introduction.html>. Visited 28.11. 2003.
- Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric statistics for the behavioural sciences*. McGraw-Hill.
- Eero Sormunen 2000. *A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases*. Acta Universitatis Tamperensis 748.
- Karen Sparck Jones 1974. Automatic indexing. *Journal of Documentation* 30(4), pp. 393 – 432.
- The Finnish stemming algorithm. Available at <http://snowball.tartarus.org/finnish/stemmer.html>. Visited 28.11. 2003