Trends in Information Management (TRIM)
ISSN: 0973-4163
9(1), pp. 1-21

Managing word form variation of text retrieval in practice – Why language technology is not the only cure for better IR performance?

Kimmo Kettunen

Abstract

Purpose: *The article discusses on a general methodological level different methods that have been used for management of single key word form variation in information retrieval during the history of textual information retrieval. The paper offers the reader an overall practical guide for choosing between different methods to be used for different types of European languages. Methods being compared in the paper include stemming, lemmatization, truncation, syllabification, unsupervised morphological methods, character n-gramming and generation of inflected word forms.*

Methodology/Approach: *Based on the empirical findings and results achieved by other researchers the paper discusses several pros and cons of different keyword variation management methods in a broader context than usually in IR, where only achieved effectiveness results are normally considered. The study proposes a list of five criteria for comparison of the conflation methods in general and offer a heuristics for choosing a suitable method for conflation of a specific language.*

Findings: *Simpler character-based methods could be preferred in IR instead of very sophisticated linguistic methods. It is also suggested that for morphologically simple languages, such as English, any kind of keyword variation management may be futile, as the increase in IR effectiveness achieved may be very low. Morphologically more complex languages can be conflated with the simple methods quite effectively for present IR search engines.*

Keywords: *Information retrieval; Management of word form variation; Comparison of word form variation management methods; IR performance; Effectiveness; Language technology*

Paper Type: *Meta-analysis*

Introduction

One of the basic problems of full-text retrieval is variation of word forms that is caused by morphology of natural languages. Shortly put, this means that one base or dictionary form of a word in language may occur in different (inflected) variant forms in texts. Out of this follows that many times the principle *"one keyword – one concept - one match"* does not hold in the textual index of retrieval systems due to morphology alone. Consequently something needs to be done to morphological variation so that the performance of information retrieval (IR) systems will not suffer too much if the language has a rich or at least medium rich morphology.

To overcome the problem of keyword variation several management methods have been proposed during the history of textual IR. The first word analysing method applied to IR was stemming, first stemmer being

Janet Lovins's stemmer for English **(Lovins 1968)**. Late 1980's saw the in-march of morphological analysis using large dictionaries, also known as lemmatization **(Alkula 2001; Koskenniemi 1996)**. During the last 10 years unsupervised morpheme detection methods **(Hammarström & Borin, 2011)** have been used somehow successfully in management of word form variation management of IR **(Kurimo, Virpioja & Turunen, 2010)**. All these methods can be characterized as reductive **(Kettunen, 2009)**: running word forms are analysed in them and reduced to either stems or base forms or morphs, if possible. The reduced forms are then used both in the indexes of search engines and as keywords in searches.

Another logical option for management of keyword variation is to use generated inflected word forms (or only inflectional stems) as search keys. In this approach, a set of inflected variant forms are generated from the input keyword and these are sought for in the plain word index of the retrieval engine. The basic fear in this method is that the language has too much inflection and too many generated word forms need to be sought for, which would make search impractical due to time considerations. But as e.g. **Kettunen and Airio (2006), Kettunen and Arvola (2012)** and **Leturia, Gurrutxaga, Areta, Alegria, and Ezeiza (2012)** have shown, only a partial generation of the most frequent inflected word forms yield good retrieval performance even for a morphologically very complex language that may in principle have thousands or even hundreds of thousands of grammatical forms.

So far mentioned methods can be characterized as linguistically motivated, either fully (morphological analysis, word form generation) or partly (stemming, unsupervised morpheme detection). A third group of methods is non-linguistic, and it includes different character string oriented methods. These include, for example, different types of keyword truncation, character n-gramming **(McNamee, Nicholas & Mayfield, 2009)** and usage of hyphen like structures **(Kettunen, McNamee & Baskaya, 2010)**. Truncation was perhaps the first method of word form variation management used in IR, and it was first based on the user's choice of proper truncation point. Lately, truncation with a fixed length (e.g. five character truncation starting from the beginning of the word) has been shown to be quite effective with many languages. N-gramming has been shown to be a language and writing system independent method **(McNamee, Nicholas & Mayfield, 2009)**.

These methods and their variants cover much of the word form variation management techniques that are actively used in IR. Reductive management techniques for word form variation are far more general than generative methods, and in practice different stemmers seem to be a standard tool of IR research. Later on we'll see, how the methods fare in relation to each other with some general criteria.

The basic problem of this article is, when to apply a certain kind of word form variation management method to a specific language. There is ample research literature concerning usage of word form variation management techniques for different languages, but the research papers are not very informative about overall benefits of the used techniques for a specific language, and mostly, only gains in recall and precision are discussed and evaluated. When one wants to set up a textual search system for a language or a group of languages, one needs to make decisions what techniques to use and what not to use. Word form variation management techniques can be very different in their scope: some, for example, are very language specific, some quite language independent. If one is setting up a multilingual textual search system, one obviously would be prone to choose a more language independent technique that handles as many languages as possible. Many other issues affect the choices, too. We'll discuss a set of choices that seem most important. In our examples we'll concentrate deliberately on a set of different European languages, while a very broad coverage of world's languages would be too ambitious a goal. We still believe that our discussion can be applied also to other languages than those discussed in the article.

Another basic theme in the article is to put IR and language technology (LT) views on morphological processing in cross-light. When word form variation management methods of IR are discussed, one needs to keep in mind, that the issue has two dimensions: that of language technology or linguistic processing and that of information retrieval. Language technology and information retrieval have partly different and partly overlapping criteria for developing and using word form processing tools. Language technology aims at linguistic felicity and as broad linguistic coverage as possible **(Koskenniemi 1996; Galvez, Moya-Anegón, & Solana 2005)**. These are justified aims, but they should be kept separate from IR performance the methods enhance. Products of LT research aid IR many times, but many times products of LT may be too sophisticated to be used when one considers the gain received in an IR context. Less LT can many times be more in IR result wise than more LT. In our opinion, information retrieval can righteously have more modest aims with its LT tools: it may be satisfied with linguistically poorer methods that improve enough effectiveness of searches in its current phase, where retrieval is based on matching of string level representations of words, not semantic entities.

The structure of the article is following: we shall first give a short account of information retrieval basics and after that we proceed to discuss different word form management methods and their relative advantages.

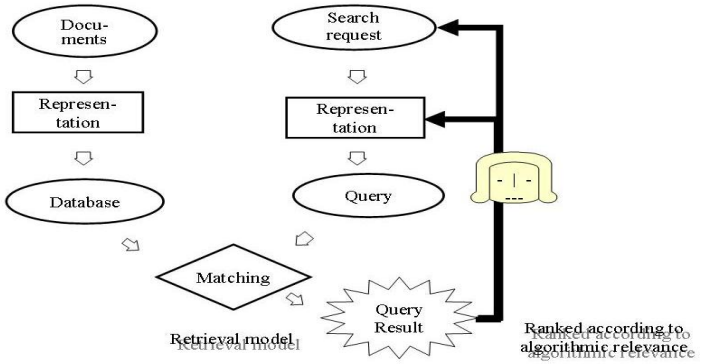
Our basic findings and recommendations are presented in section 3, and section 4 draws some more conclusions on the issue.

IR basics

For discussion we need to outline first working principles of a state-of-the-art text search engine. The description of IR is based on two current textbooks, **Croft, Metzler and Strohman (2010)** and **Ingwersen and Järvelin (2005).** Due to space requirements the discussion is very concise and basic, and an interested reader is asked to look after the references and other IR sources for further details.

By a text information retrieval system we mean a textual database system consisting of text documents and means to manage the database. Documents in the database can be searched for, and new documents can be added to the database if needed. Searching in the textual database is based on matching of a query term representation and an inverted index that represents the contents of the documents as index terms. Another important feature of IR systems is *ranking*: returned documents are given as an ordered list where the documents expected to be most relevant are at the top and less relevant in decreasing order of relevance. **(Croft, Metzler, & Strohman, 2010 a; Ingwersen and Järvelin 2005, a)**. **Figure 1**. gives an outline of the overall situation including the search engine user.

Fig. 1: A simplified picture of an IR system, adapted from Ingwersen and Järvelin (2005, b)



The basic goal for an IR engine is to fulfil user's information need as well as possible. The more the engine returns relevant documents at the top of the result list, the better it is. Users, however, may be satisfied with only a few highly relevant documents at the beginning of the result list. This is especially true with web searches **(Ingwersen & Järvelin 2005)**.

Management of word form variation in an IR engine *may* help in achieving this goal. This is especially true with non-Web search engines, but with Web search engines the role of word form variation management is more diffuse, as there are many other interfering factors, such as effect of links etc. A recent study by **Uyar (2009)** shows, that Google's handling of English, a morphologically simple language, is quite complex. Word forms are sometimes stemmed to the index, sometimes not, depending on the documents, and principles behind this seem to be diverse. Many non-English languages will most probably be handled differently from this in Web search engines **(Lazarinis, Vilares & Efthimiadis, 2009)**.

Why and when should word form variation management be used in IR?

- **Languages are different**

As shown in the IR basics part, queries and documents are matched in the IR database according to their string level representations. Singular and plural surface forms of lexeme {*cat}*, *cat* and *cats*, do not match, if something, like stemming, is not done to them to make the representations similar. In all of the word form variation management methods the basic principle is the same: decreasing of variation found in natural language word forms. Amount of variation in different languages differs very much: English nouns may have four forms, Finnish nouns may in principle have about 2000 grammatical forms, Basque about 458,683 **(Leturia, Gurrutxaga, Areta, Alegria & Ezeiza 2012)**. It is clear that different languages set different needs for word form variation management used in an IR engine.[1]

The explanation for varying behaviour of words in different languages is linguistic complexity. On morphological level linguistic complexity means roughly, that the language has lots of inflection, which is realized, for example, in number of different nominal case forms the language has **(e.g. Iggesen 2011)**. Finnish, for example, has 14 different cases, and English has two. This means that Finnish has many variating word forms, as English has few due to the nature of the case morphology. The number of different possible forms for a basic inflected noun – without clitics or

[1] *We assume here, that what matters in IR as keywords, are mainly nouns. This can be easily seen from basic linguistic knowledge: nouns are the largest open word form class in any language, and majority of words in texts or dictionaries are nouns. Nouns denote to entities of the world, real or invented or possible, and other word classes either modify them (as adjectives: a big dog) or relate them to each other in sentences (verbs: The dog sleeps on the floor). There may be slight language dependent differences with this, but we consider this a solid basic principle. Adjectives as modifiers seem to have some effect for search results in our experience. But their effect in the MAP of morphologically more complex languages such as Finnish, Swedish and German, is about 1-2 per cent units in the MAP. The effect of the main word classes (nouns, adjectives and verbs) to IR results should of course be tested empirically with a large set of languages word class by word class.*

possessive endings - of the EU languages varies from two (Dutch, Spanish and other Romance languages) to about 40 (Hungarian). Many times already the number of cases in the language is indicative of morphological complexity, but not always (e.g. in the case of Swedish and Bulgarian). Then other morphological categories, such as marking of definiteness and expression of number in the language, are the key factors **(Stump, 2001)**. Compounding, creation of new words by concatenating existing words to form new ones, gives some added complexity to some languages, at least in the IR context.

The *'IR hardness'* of a language is clearly related to its morphological complexity. **McNamee, Nicholas & Mayfield (2009, Table 6)** show this by relating length of words (the longer the words in the language, the more morphemes they have), two linguistic complexity ratios and gains in IR performance achieved with 5-grams. These figures correlate at least moderately (lowest correlation being 0.68) or very highly (highest correlation being 0.91). **Kettunen (2009)** shows the same informally by counting the difference of best and worst mean average precision (MAP) results of IR performance for the language. The bigger the difference, the more morphologically complex the language is. Same kind of measurement idea is **Pirkola's (2001)** suggestion of using language typological information as an indication of need for word form variation management for a certain type of language. Pirkola's suggestion is based on the indices of synthesis and fusion. Index of synthesis "*refers to the amount of affixation in a language, i.e. it shows the average number of morphemes per word in a language*". Index of fusion "*refers to the ease with which morphemes can be separated from other morphemes in a word. Agglutinative languages have a low index of fusion, and in fusional languages it is high. In agglutinative words segmentation can be performed readily owing to clear morpheme boundaries. In fusional words segmentation is difficult or impossible*". Both of these indices are scales, where different languages are between the two opposite poles. Pirkola suggests that typological information of languages could be readily applied to linguistic problems of keyword and index term variation and stemmer creation, but his suggestions lack realistic large scale data in the article.

In our opinion studies of morphological complexity of different languages offer a more realistic basis for applying generalized linguistic knowledge to IR. Approximation of the morphological complexity of a language can be operationalized readily with text corpora. Following we describe a few approaches.

A popular way to approximate the morphological complexity of a language has been **Juola's (1998, 2008)** suggestion of distorting the word structures by using random numbers for each different word type. After

distortion, the data is compressed using a compression algorithm. Then the size of the compressed original word file is divided by the size of the compressed distorted word data file. The result tells the complexity of each language's morphology on the basis of *Kolmogorov complexity* that the compression algorithm approximates. The method has been used besides Juola by **Sadeniemi, Kettunen, Lindh-Knuutila & Honkela (2008)** and **Ehret and Szmrecsanyi (2011)** for a set of languages. Ehret and Szmrecsanyi show that the Juola method suits equally well to parallel, semi-parallel and non-parallel texts, the last being a very good addition to the method's applicability. Besides they show that the method is robust with respect to different sampling points in the process.

Bane (2008) suggests another kind of approach for approximation of morphological complexity of languages. He uses *Linguistica*, software that induces morphology of a language from a given text sample. Based on the analysis of an un-annotated text corpus, Linguistica separates word stems, affixes and signatures. In this approach the morphological complexity of the language is based on the formula

$$\text{Morphological Complexity} = \frac{\text{DL(Affixes)} + \text{DL (Signatures)}}{\text{DL (Affixes)} + \text{DL(Signatures)} + \text{DL (Stems}}$$

Affixes and stems in the formula describe linguistic units identified by the Linguistica software, signatures describe the "possible distributions of affixes upon stems". Thus, the method measures "a language's morphological complexity as the proportion of the lexicon's total description length that is due to the description lengths of the affixes and signatures" **(Bane 2008)** [2]

It should, however, be noted that all the morphological complexity figures give approximations for whole of the morphology of each language. We suggest that for IR purposes the complexity figures could be counted for noun corpuses only (they could include adjectives, as well). Anyhow, even the current complexity figures seem to correlate well with achieved MAPs, as we'll see in more detail later.

Also one simpler calculation can be used in complexity assessment. **Juola (1998)** shows that type-token relations of word forms in a language indicate also the morphological complexity of a language. Morphologically simpler languages use fewer word types and more tokens, and morphologically more complex languages use clearly more types and fewer tokens. This information could be used in assessing the

[2] *Bane gives examples with 19 translations of the Bible. Most of the figures given by Bane seem realistic, but e.g. Italian, Spanish and French seem to be quite high in the listing, thus approximated more complex than e.g. Swedish and German.*

morphological complexity of a language as well. It is easily countable and corpora are readily available.ll

- **How different languages behave in IR**

In this section we show how different languages behave in an information retrieval context. We'll show results of two studies that have either direct empirical results from IR evaluation of several languages and word form variation management methods or have collected such data from other studies. These empirical IR results are related to morphological complexity of the languages.

McNamee, Nicholas & Mayfield (2009) use 18 different methods for management of word form variation for 18 languages in five different writing systems and of different morphological complexity. Methods for word form variation management include all the main methods used in IR, except lemmatization that is not easily available for such a variety of languages. Instead of lemmatization, two types of stemmers, rule-based and statistical, are used. Different phonetic transformations (Soundex and devowelization), truncations and character gramming (n-gramming, where n varies from 3 to 7, and skip-gramming, where some of the characters may be skipped) are included in the methods. No generative methods are present in *McNamee et al's (2009)* study.

The main results of the paper are the following:

- character n-gramming is the most effective method for most of the languages, the length of N being four or five characters
- rule based stemming (Snowball stemmers are used) can be an attractive option for languages where morphological variation is not very high
- phonetic transformations do not work well for any language (and thus they can be forgotten from the repertoire)
- a statistical stemmer (i.e. a particular unsupervised morphological method, Morfessor) does not perform too well, but is getting better (cf. also **Kurimo, Virpioja, & Turunen (2010)** for the latest results with different unsupervised morphological systems)
- one of the most unsophisticated and un-linguistic methods, five character truncation, works very well with most of the languages, being the second best non n-gram method overall, only slightly behind performance of Snowball stemmers.

Table 1 combines parts of results of **McNamee, Nicholas & Mayfield (2009)** and collected IR data from **Kettunen (2009)**, and shows the situation with 14 languages that have available IR collections and data. Many morphologically interesting European languages, such as Estonian, Latvian and Lithuanian, are unfortunately missing from the table, as there are no IR collections for these languages, but the variation in languages is enough to make our points.

Column two and three in the table show basically the same thing, difference between the IR result when best possible available word form variation management method has been used for the language versus situation when plain word forms have been used. The information in columns two and three shows the bounds of performance improvement gained with word form variation management for each language. As the figures show, results from different collections, search engines and word form variation management methods are quite similar. Column four interprets need for word form variation management according to **Sparck-Jones's (1974)** old rule: if the statistically significant absolute difference in MAP is under 5 %, the practical difference is not noticeable (for the user); if the MAP difference is over 5 % but under 10 %, the practical difference is noticeable. When the difference is over 10 %, the practical difference is material. These percentage figures are stated here as *no need*, *beneficial* and *necessary* in the table.

Table 1: Necessity of word form variation management in the light of MAP results.

	Language	GAP = best MAP with word form variation management *minus* plain words MAP (Kettunen 2009)	Lowest and highest MAPs gained (McNamee, Nicholas & Mayfield 2009) low	high	Is word form variation management needed for the language?	Morphological complexity figure (Sadeniemi, Kettunen, Lindh-Knuutila, & Honkela, 2008)
1	Bulgarian	6.8-8.1 %	0.216	0.31	beneficial	N/A
2	**Czech**	N/A	0.227	0.329	necessary	1.0867
3	Dutch	0.6.-5.0 %	0.381	0.424	no need	1.1189
4	English	1.2-2.9 %	0.406	0.437	no need	1.0529
5	**Finnish**	10.5-25.2 %	0.34	0.507	necessary	1.1637
6	French	0.5-3.8 %	0.363	0.401	no need	1.0622
7	**German**	6-15.7 %	0.33	0.42	beneficial/ necessary	1.1660
8	**Hungarian**	9.9-12.4 %	0.197	0.374	necessary	1.1421
9	Italian	N/A	0.374	0.417	no need	1.0518
10	Portuguese	N/A	0.316	0.352	no need	1.0676
11	**Russian**	6.1-21.0 %	0.267	0.373	necessary	N/A
12	Spanish	N/A	0.439	0.484	no need/ beneficiary	1.0624
13	Swedish	1.7-8.8 %	0.338	0.427	beneficial	1.125
14	**Turkish**	12.3 %	N/A [3]		necessary	N/A

Note: Morphologically most complex languages in the light of MAP results are in bold

[3] *McNamee et al. do not have Turkish in their repertoire, but empirical results of Kettunen et al. (2010) confirm GAP result shown in Kettunen (2009) with several word form management methods.*

In some cases (Bulgarian, German and Swedish), the line between *beneficial* and *necessary* is quite narrow, and in most of the cases of *no need,* there is no question of the borderline. Only Spanish seems to be close the 5 per cent edge. Mapping of Sparck-Jones's three part classification with mean average precision results divides the 14 languages to three groups. Those language belonging to the most complex necessary group, are Czech, Finnish, German, Hungarian, Russian and Turkish. Those in the beneficial group are Bulgarian and Swedish. Rest of the languages, Dutch, Italian, French, Italian, Portuguese and Spanish are in the no need group. We shall suggest later on how this partition could be put in practical use in IR.

If we consider the morphological complexity ratios in column five of Table 1, most of them are in accordance with the achieved absolute MAP increase for the language. The more the MAP for the language increases with word form variation management, the bigger the morphological complexity figure for the language is and vice versa. The only exception is Dutch: it does not gain much from word form variation management, but its morphological complexity figure is relatively high. We believe that here shows the discrepancy between morphological complexity of nouns and morphological complexity in general. Dutch has simple noun morphology with no cases for nouns, but it also has compounds, which might in part increase the complexity figure.

Table 2: Morphological complexity and maximal MAP increase

	Morphological complexity	Maximal reported gain in MAP, absolute %
DE	1,166	15,7
FI	1,1637	25
HU	1,1421	12,4
SV	1,125	8,8
NL	1,118	5
CZ	1,0867	10,2
PT	1,0676	3,63
ES	1,0624	4,5
FR	1,062	3,8
EN	1,052	2,9
IT	1,0518	4,29
		Correlation 0.85

In Table 2 we show that complexity figures for languages and increases in MAPs correlate highly (0.85, using *Excel's standard correlation co-efficiency formula*). The data in the MAP increase columns has been taken from different publications, and is the same as in **Kettunen (2009),** except for Czech, Portuguese, Italian and Spanish (bolded in Table 2), which originate from **McNamee, Nicholas and Mayfield (2009)**. The table is

sorted in descending order of morphological complexity, and the languages seem to be quite well ordered also with respect to their MAP increases. Only Czech should be above Swedish and Dutch in this respect. Difference of German and Finnish in MAP figures is high, although they are on a par with respect to the morphological complexity figure.
If we correlate plain word MAPs from **McNamee, Nicholas & Mayfield (2009)** to the linguistic complexity figures, there is a medium negative correlation. Thus, it seems that increase achieved with word form variation management correlates with the morphological complexity of the language, but the achieved basic level MAP does not. Figures are shown in Table 3.

Table 3: Morphological complexity and plain word form MAPs

	Morphological complexity	Maximal reported gain in MAP, absolute %
DE	1,166	0,3303
FI	1,1637	0,3406
HU	1,1421	0,1976
SV	1,125	0,3387
NL	1,118	0,3813
CZ	1,0867	0,227
PT	1,0676	0,3162
ES	1,0624	0,4396
FR	1,062	0,3638
EN	1,052	0,406
IT	1,0518	0,3749
		Correlation -0.41

- **What criteria to use for choosing a word form variation management method?**

Many methods of word form variation management of IR work considerably well from the viewpoint of effectiveness, which is measured in precision and recall (P/R) of retrieval using different measures, one of the most used being mean average precision, MAP, in Table 1. The methods can also be compared on a more general level. Three kinds of benefits are usually associated with different types of keyword variation management in IR according to **Harman (1991)**. They are briefly as follows:

- **What criteria to use for choosing a word form variation management method?**

Many methods of word form variation management of IR work considerably well from the viewpoint of effectiveness, which is measured in precision and recall (P/R) of retrieval using different measures, one of the most used being mean average precision, MAP, in Table 1. The

methods can also be compared on a more general level. Three kinds of benefits are usually associated with different types of keyword variation management in IR according to **Harman (1991)**. They are briefly as follows:

- ease of use (morphology of query words is taken care of by the retrieval system),
- storage savings - the index compression factor, ie. smaller indexes when for example lemmatization or stemming is used **(Galvez, Moya-Anegón, & Solana 2005)**, and
- improved retrieval performance.

Besides these criteria, there are, however, others that should be taken into consideration. Linguistic methods of word form variation management use many times lexicons in their analysis, and thus the lexical coverage of the morphological method used is important. This is an issue that affects lemmatizers and stemmers using dictionaries. Their dictionaries lack words for many reasons, and one of the main classes of lacking words are different kinds of proper names (persons, companies, geographical names etc.), which are usually an important subclass of query words **(Pirkola & Järvelin 2001)**. A statistical lemmatizer, such as e.g. Stale **(Loponen & Järvelin 2010)**, in turn, does not suffer from this hinder, and performs also competitively with a lexical lemmatizer in an IR context. Word form generators can be implemented without lexicons, and thus they avoid the problem of lexical coverage.

Other more technical criteria can also be used for comparison. **Croft, Metzler & Strohman (2010 b)** list the following: elapsed indexing time, indexing processor time, query throughput, query latency, indexing temporary space and index size. These criteria are related to search engine efficiency and are especially important when commercial search engines are developed and used.

We have chosen to Table 4 five different evaluation criteria for word form variation management methods used in IR. The criteria are language independence of the method, its IR effectiveness, size of the retrieval indexes created with the method, ease of rule generation for the management method and overall simplicity of the approach.

Language independence of the method means that the same algorithm and resources can be applied to many languages without modifications. N-gramming is a good example of total language independence, it suits any alphabetical language. Syllabification can be considered partly language independent: one syllable rule suits many languages **(cf. Kettunen, McNamee & Baskaya, 2010)**, but is not optimal for all. A lemmatizer using lexicons is not language independent: its lexicons and morphological rules have to be described for each language anew.

IR effectiveness is easy to define with relation to recall and precision increase the method causes. Size of the retrieval indexes is also easy to show: e.g. five character truncation leads to very small indexes, n-gramming to huge ones. Ease of rule creation refers to human resources needed in making of the linguistic system. Some can be created totally automatically or semi-automatically, some, like stemmers and lexical lemmatizers need quite much human labour. Overall simplicity of the approach tries to generalize the whole of the approach. [4]

These criteria are by no means exhaustive and also others could be included or some omitted. Efficiency considerations have been left out of our criteria, because there is no available data related to them and efficiency is also so dependent on a specific implementation.

The methods have been assessed with pluses +++, ++ and +. With 0 the effect is not positive or not applicable, with + weak, with ++ effect is clearly positive, mid-size, and with +++ there is a big positive effect, or best performance.

A short description of scoring of the methods in the Table 3 is in order, criterion wise. Highest language independence of automatic truncation, unsupervised morphological methods, n-gramming and plain words is clear. Application of truncation depends on the language's way of placing morphological affixes to words. In the case of a clear suffixing language truncation from the end works. If the language is prefix oriented, truncation from the beginning should work. For infix oriented languages **McNamee, Nicholas & Mayfield's (2009)** least frequent substring would probably be the best approach. Lemmatizers and stemmers need to be crafted language by language, if they are not based on statistical knowledge, and thus they are weakly language independent. Word form generation has been given one plus, while it seems to have one language independent feature: distributions of inflected forms follow a statistical principle that can be utilized in search applications.

The IR effectiveness of each method can be assessed most objectively of all the criteria. Lemmatizers cause most increase in MAPs in most of the evaluations for most of the languages, if they are used. All other methods except plain words cause a medium sized increase.

Index size effects have been discussed astonishingly much for example in the stemmer literature. However, the real effect of the size of an index is a bit obscure for most of the methods. Only in the case of n-gramming the indexes grow obviously so much **(McNamee, Nicholas & Mayfield, 2009; Table 5)** that usage of n-gramming is most probably counterproductive in a real-time search engine.

[4] *Harman's (1991) ease of use is omitted from the table, as it is actually included in all of the methods.*

Table 4: Scoring of different word form variation management methods along five criteria

Method and reference to an example	Language independence	Effectiveness	Index size	Ease of rule generation	Simplicity of the approach	SUM
automatic truncation **(McNamee, Nicholas & Mayfield, 2010)**	+++	++	+++	+++	+++	**14**
unsupervised morphological methods **(Hammarström and Borin 2011; Kurimo, Virpioja & Turunen, Nicholas, C. & Mayfield, J. (2009 2010)**	+++	++	++	+++	++	**12**
Syllabification **(Kettunen, McNamee & Baskaya, 2010)**	++	++	+	++	+++	**10**
n-gramming (plain, no skips) **(McNamee, Nicholas & Mayfield ,2009)**	+++	++	0	+++	++	**10**
statistical lemmatization **(Loponen and Järvelin 2010)**	++	++	++	++	++	**10**
plain words	+++	0	+	+++	+++	**10**
lemmatization (rules + a dictionary) **(Koskenniemi 1996)**	0	+++	++	0	+	**6**
rule based stemming **(Snowball web site)**	0	++	++	0	++	**6**
word form generation **(Kettunen and Airio, 2006; Leturia, Gurrutxaga, Areta, Alegria & Ezeiza, 2012)**	+	++	+	0	++	**6**

In the criteria *ease of rule generation* two methods need a comment. We have given syllabification and statistical lemmatization two pluses. In the case of syllabification, based on empirical results of 14 languages **(Kettunen, McNamee & Baskaya, 2010)** it seems that even one simple rule is quite effective for a variety of languages although it is not optimal for some of these. Statistical lemmatization in the style of Stale **(Loponen and Järvelin 2010)** needs first setting up of a learning corpus. But this phase can be quite easy to do.

The case of overall simplicity of the approach is the most diffuse. It involves the whole delivery chain, so to speak: from production of a word form management method to its attachment to the query engine. A rule and dictionary-based lemmatizer is given the lowest score, while its lexicons need constant updating and search engine indices constant lemmatizing. Rule-based stemming, statistical lemmatization, n-gramming, unsupervised morphological methods and generation are considered here more simple. Even if an inflected query word generator may need quite a lot work in setting up, it is easily attached to a query engine through only a search API, which gives it a great plus **(Leturia, Gurrutxaga, Areta, Alegria & Ezeiza, 2012)**. Plain words need nothing special, and truncation and syllabification are also simple from an overall perspective: they are easy to produce and quite easy to apply in the search engine.

When figures of the Table 4 are examined, we can see that simpler character oriented methods get more pluses. Five character truncation and unsupervised morphological methods are the two best methods here, in this order. After them come syllabification, n-gramming, statistical lemmatization and plain words on a tie. Rule-based stemming and lemmatization with rules and a dictionary do not fare too well, although they are the two most used methods of word form variation management in IR research.

The results and the chosen assessment criteria are of course open to discussion, but in our opinion they do reflect important details that should be taken into consideration when choosing word form variation management method for an IR system. A list of this type should help in making decisions when different techniques are considered for use in an IR engine. Weights for different criteria can be use-case dependent, too, and one could weight some of the criteria more depending on the case. For example, if IR effectiveness is the most important thing required from the system, then effectiveness should be over-weighted. In a multilingual document collection one would probably overweight language independence of the LT method.

- **A heuristics for use of word form variation management methods**

So far we have been giving the basic cornerstones for practical recommendations to be given. In this section practical recommendations for usage of word form variation methods in IR of different languages are given.

Based on the data in **Tables 1** and **4** a heuristic recommendation for usage of different word form variation management methods in IR would be as follows – the heuristics applies for other languages not shown here, too. The heuristics follows the grouping of languages in Table 1 and consists of three general initial considerations and three language and word form variation method oriented points.

Our *first general heuristic* rule for use of word form variation management with any language is this: consider the morphological complexity of the languages(s) that need to be handled in the search engine. This can be done by using the methods we have discussed in **section 3.1**. One can either:

- Consult IR results for the language, if they are available. The difference of the best and worst MAP (GAP in **Kettunen, 2009**) will show approximately what the morphological complexity of the language is **and** what can be gained by using word form variation management at best.
- If no IR results for the language are available, morphological complexity of the language can be assessed e.g. with the **Juola (1998)** method reliably enough. Word form data for assessment should be usually easily available.
- Also linguistic literature can be used in assessment. Counting of the number of noun cases is already a very good approximation of the morphological complexity of a language. Type-token ratios for words are also useful.

From this phase you'll gain insight, what kind of improvements in IR performance you can expect from word form variation management for a specific language. This will also envisage you in the choice of a specific word form variation management method.

Our *second general heuristic* rule is this: If there are several languages that need to be handled for the search engine, a suitable method that applies for all or most of the languages should be given precedence over methods that apply for a single language only. This simplifies both implementation and maintenance of the IR system.

The *third general heuristic* rule is obvious: check what kind of word form variation management programs are readily available for the language(s) you need to retrieve in the search system. For well-resourced languages a varying repertoire is possibly available freely, rest of the available are proprietary and need to be paid for. For small and less-resourced

languages the choice of available programs may be restricted. If you consider implementing something like an algorithmic stemmer yourself – this is not a huge task – this is also feasible, but before doing so, you should consider its necessity first in the light of the following heuristics.
Following three heuristics can be used to figure out, what kind of word variation management could be used after the initial considerations.

For morphologically simple languages - group *not necessary*, such as 3, 4, 6, 9 in Table 1 - do nothing but normal routines (case folding, tokenization etc.). Plain word forms are a good solution for indexing and query formation with these languages. There is not much to be gained anyhow IR wise with linguistic means, so it is not necessary to use any word form variation management methods with these languages. If you really want to do something with these languages, choose the simplest methods possible.

If a morphologically simple language is a compounding language that merges together words to create new words (such as for example Dutch is), a splitting procedure for compounds together with truncation might be a better solution than truncation alone.

1. If the language is in the *beneficial* group - such as 1, 7, and 13 in Table 1 - the simplest non-linguistic word form management method can still be used quite well. Out of the simple methods five character truncation is the easiest to implement and very effective, but also n-gramming and hyphenation could be used. Large indexes and slow retrieval are shortcomings of n-gramming, so if your search application is time critical, n-gramming is not a good choice. A light stemmer can also be considered, if such is available or can be easily implemented. But there is no need for 'heavy artillery' here, either.
2. With languages in the *necessary* group - such as 2, 5, 8, 11 and 14 in Table 1- one can begin to seriously consider also 'heavier' methods, such as stemming or lemmatization, as there is really something to be gained IR result wise. Even here they are not necessary, as five character truncation is relatively effective with these languages too. If one's only need is to have good IR performance from the search engine, then language technology oriented tools may be overkill. If one has also other needs for the linguistic analysis capabilities of the IR system (such as handling of lemmas or interaction as e.g. in query expansion **(cf. Galvez, Moya-Anegón & Solana 2005)**, then one may consider an elaborate lemmatizer.

Conclusion

The paper discusses and compares usefulness of different word form variation management methods for IR and given practical suggestions for choosing the methods. The issue is far from simple, and many arguments

can be given pro different solutions. A low level approach has been deliberately taken, where need of very high level morphological tools with IR has been partly questioned. The study also shows connections of morphological complexity and IR performance improvements with word form variation management and suggested how this information can be used in practice. Finally, a heuristics for choosing different word form variation management methods to be used with IR is suggested.

The considerations and suggestions of the paper come near to Ken Church's DDI claim (Don't Do It), which states that morphology aware software should perhaps not be used at all in computational handling of language: "*There are lots of morphology programs out there, many of which work surprisingly well. Nevertheless, for many practical applications, we prefer not to use such programs, if we have the choice. Simple morphological inferences are better than complex inferences. But even simple inferences are worse than none*" **(Church 2005).** When examining the data, this seems partly true with regard to the role of morphology programs in IR: you can skip proper morphological processing with use of simple string manipulation and get good results anyhow. Sometime all morphological inferences can be skipped (cf. Table 1, no need languages), and most of the times simple inferences do the trick.

Another, more theoretical, argument in favour of simpler methods is Minimal Description Length (MDL), which basically formalizes the old Occam's razor: when two models fit the data equally well, MDL will choose the one that is the simplest in the sense that it allows for a shorter description of the data **(Grünwald 2007)**. If we apply the idea of MDL for morphological components used in IR, we can e.g. say that five character truncation could be favoured instead of a lemmatizer, as it is far simpler and "fits the data" – i.e. management of word form variation for IR – almost as well as stemming or lemmatization with many languages. A five character truncation module for a search engine can be coded in about two to three code lines in almost any programming language, when a lexical lemmatizer needs description of lexicons (tens of thousands of lines) and a rule component (a few hundred lines). The same argument would apply for simple syllabification, although it is slightly more complex on the index side representation. Other methods are between these extremes.

Acknowledgements

This paper was finished while the author was visiting UFAM, Universidad Federal do Amazonas, Institute of Computing, and funded by FAPEAM, Fundação de Amparo à Pesquisa do Estado do Amazonas (http://www.fapeam.am.gov.br/)

I wish to thank Professor Kalervo Järvelin and Dr. Heikki Keskustalo, both from the University of Tampere, School of Information Sciences, for commenting earlier versions of the article.

References

Alkula, R. (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 (3-4), 195−208.

Bane, M. (2008). Quantifying and measuring morphological complexity. In *Proceedings of the 26th West Coast Conference on Formal Linguistics* (pp. 67–76). Retrieved from http://www.lingref.com/cpp/wccfl/26/paper1657.pdf

Church, K.W. (2005). The DDI approach to morphology. In A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund and A. Yli-Jyrä (eds.), Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday. (p. 25-34). Retrieved from http://cslipublications.stanford.edu/koskenniemi-festschrift/kk-festschrift-all-2005.pdf

Croft, B. W., Metzler, D. & Strohman, T. (2010). *Search Engines. Information Retrieval in Practice.* Boston, Paris: Pearson.

Croft, B. W., Metzler, D. & Strohman, T. (2010 a). *Search Engines. Information Retrieval in Practice (pp. 13-28).* Boston, Paris: Pearson.

Croft, B. W., Metzler, D. & Strohman, T. (2010 b). *Search Engines. Information Retrieval in Practice (p. 327).* Boston, Paris: Pearson.

Ehret, K. & Szmrecsanyi, B. (2011). *An information-theoretic approach to assess linguistic complexity*. Retrieved from http://www.benszm.net/omnibuslit/EhretSzmrecsanyi_web.pdf

Galvez, C., Moya-Anegón, F. de & Solana, V. H. (2005). Term conflation methods in information retrieval. Non-linguistic and linguistic approaches. *Journal of Documentation,* 61 (4), 520–547.

Grünwald, P. (2007). *The Minimum Description Length Principle (p.29)*. Cambridge, Mass: MIT Press.

Hammarström, H. & Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics,* 37 (2), 309–350.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science,* 42 (1), 7-15.

Iggesen, O.A. (2011). Number of cases. In M. S. Dryer M. and Haspelmath (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 49A. Retrieved from http://wals.info/chapter/49A

Ingwersen, P. & Järvelin, K. (2005). *The Turn. Integration of Information Seeking and Retrieval in Context.* Dordrecht : Springer.

Ingwersen, P. & Järvelin, K. (2005 a). *The Turn. Integration of Information Seeking and Retrieval in Context (p. 119).* Dordrecht : Springer.

Ingwersen, P. & Järvelin, K. (2005 b). *The Turn. Integration of Information Seeking and Retrieval in Context (p.115).* Dordrecht : Springer.

Juola, P. (1998). Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics*, 5(3), 206–13.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki and F. Karlsson (eds.) *Language Complexity : Typology, Contact, Change*. Amsterdam: John Benjamins Press.

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

Kettunen, K. & Arvola, P. (2012). Generating variant keyword forms for a morphologically complex language leads to successful information retrieval with Finnish. In B. Larsen and M. Salampasis (eds.), Advances in Multidisciplinary Retrieval, 5th Information Retrieval Facility Conference (pp. 113-126).

Kettunen, K. (2009). Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval. *Journal of Documentation,* 65 (2), 267–290.

Kettunen, K., McNamee, P. & Baskaya, F. (2010). Using syllables as indexing terms in full-text retrieval. In I. Skadina, A. Vasiljevs (eds), *Human Language Technologies, the Baltic Perspective* (pp. 225−232*)*. IOS Press.

Koskenniemi, K. (1996). Finite state morphology and information retrieval. *Natural Language Engineering,* 2 (4), 331–336.

Kurimo, M., Virpioja, S. & Turunen, V. (eds.) (2010). Proceedings of the Morpho Challenge 2010 workshop. Technical Report TKK-ICS-R37, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland. Retrieved from http://research.ics.aalto.fi/events/morphochallenge2010/papers/ProcMorphoChallenge2010.pdf.

Lazarinis, F., Vilares, J., Tait, J. & Efthimiadis, E. (2009). Current research issues and trends in non-English Web searching. *Information Retrieval,* 12 (3), 230-250.

Leturia, I., Gurrutxaga, A., Areta, N., Alegria, I. & Ezeiza, A. (2012). Morphological query expansion and language-filtering words for improving Basque web retrieval. *Language Resources & Evaluation*. DOI: 10.1007/s10579-012-9208-x

Loponen, A. & Järvelin, K. (2010). A dictionary- and corpus-independent statistical lemmatizer for information retrieval in low resource languages. In M. Agosti, N. Ferro, C. Peters, M. de Rijke, and A.

Smeaton (eds.) CLEF'10 Proceedings of the 2010 international conference on Multilingual and multimodal information access evaluation: cross-language evaluation forum. LNCS vol. 6360, (pp. 3-14). Heidelberg: Springer.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistic,* 11, pp. 23–31.

McNamee, P., Nicholas, C. & Mayfield, J. (2009). Addressing morphological variation in alphabetic languages. In Proceedings of the 32nd Annual International Conference on Research and Development in Information Retrieval (SIGIR-2009), Boston, MA, 75-82.

Pirkola, A. & Järvelin, K. Employing the resolution power of search keys. *Journal of the American Society for Information Science and Technology*, 52 (7), 575−583.

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation* 57 (3), 330-348.

Sadeniemi, M., Kettunen, K., Lindh-Knuutila, T. & Honkela, T. (2008). Complexity of European Union languages: a comparative approach. *Journal of Quantitative Linguistics*, 15 (2), 185–211.

Sparck-Jones, K. (1974). Automatic indexing. *Journal of Documentation,* 30 (4), 393-432.

Stump, G. T. (2001). Inflection. In A. Spencer and A. Zwicky (eds), *The Handbook of Morphology* (pp.13-43). Hoboken, NJ: John Wiley and Sons.

Uyar, A. (2009). Google stemming mechanisms. *Journal of Information Science,* 35 (5), 499-514.

Corresponding Author

Kimmo Kettunen can be contacted at: Kimmo.Kettunen@uta.fi

Author Biography

Dr. Kimmo Kettunen finished his Ph.D. in information retrieval at the University of Tampere, Finland, in 2007. His main areas of research have been mono- and cross-lingual IR, machine translation, and language technology. He spent winter and spring of year 2013 as a visiting professor at the Universidade Federal do Amazonas, Institute of Computing, in Manaus, Brazil.