Corpus-based Paradigm Selection for Morphological Entries

Krister Lindén
University of Helsinki
Helsinki, Finland
Krister.Linden@helsinki.fi

Jussi Tuovila
University of Helsinki
Helsinki, Finland
Jussi.Tuovila@helsinki.fi

Abstract

Language software applications encounter new words, e.g., acronyms, technical terminology, loan words, names or compounds of such words. To add new words to a lexicon, we need to indicate their inflectional paradigm. In this article, we evaluate a lexicon-based method augmented with data from a corpus or the internet for selecting the inflectional paradigm of new words in Finnish. As an entry generator often produces numerous suggestions, it is important that the best suggestions be among the first few, otherwise it may become more efficient to create the entries by hand. By generating paradigm suggestions with an entry guesser and then further generating key word forms for the suggested paradigms, we were able to find support for the paradigms in a corpus. Our method has 79-83 % precision and 86-88 % recall, i.e. an F-score of 83-86 %, i.e. the first correctly generated entry is on the average found as the first or the second candidate.

1 Introduction

New words are constantly finding their way into daily language use. This is particularly prominent in rapidly developing domains such as biomedicine and technology. The new words are typically acronyms, technical terminology, loan words, names or compounds of such words. They are likely to be unknown by most hand-made morphological analyzers. In many applications, hand-made guessers are used for covering the low-frequency vocabulary or the strings are simply added as such.

Mikheev (1996, 1997) pointed out that words unknown to the lexicon present a substantial problem for part-of-speech tagging, and he presented a very effective supervised method for inducing English guessers from a lexicon and an independent training corpus. Oflazer & al. (2001) presented an interactive method for learning morphologies and pointed out that an important issue in the wholesale acquisition of open-class items is that of determining which paradigm a given citation form belongs to.

Recently, unsupervised acquisition of morphologies from scratch has been studied as a general problem of morphology induction in order to automate the morphology building procedure. For overviews, see Wicentowski (2002) and Goldsmith (2007). If we do not need a full analysis, but only wish to segment the words into morph-like units, we can use segmentation methods like Morfessor (Creutz & al., 2007). For a comparison of some recent successful segmentation methods, see the Morpho Challenge (Kurimo & al., 2007).

Although unsupervised methods have some advantages for less-studied languages, for the well-established languages, we have access to fair amounts of lexical training material in the form of analyses in the context of more frequent words. Especially for Germanic and Fenno-Ugric languages, there are already large-vocabulary descriptions available and new words tend to be compounds of acronyms and loan words with existing words. In English, compound words are written separately or the junction is indicated with a hyphen, but in other Germanic languages and in the Fenno-Ugric languages, there is usually no word boundary indicator within the compounds. It has previously been demonstrated by Lindén (2008) that already training sets as small as 5000 inflected word forms and their manually determined base forms will give a reasonable result for guessing base forms of new words by analogy, which was tested on a set of languages from different language families, i.e. English, Finnish, Swedish and Swahili.

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 96–102

In addition, there are a host of large but shallow hand-made morphological descriptions available, e.g., the Ispell collection of dictionaries (Kuenning, 2007) for spell-checking purposes, and many well-documented morphological analyzers are commercially available, e.g. Lingsoft[1]. It has also been demonstrated by Lindén (2009) that there is a simple but efficient way to derive an entry generator from a full-scale morphological analyzer implemented as a finite-state transducer. Such an entry generator can be used as a baseline for more advanced entry guessing methods.

In this work, we propose and evaluate a new method for *selecting the inflectional paradigm for an inflected word form* of a new word by generating paradigm suggestions with an entry generator and then further generating key words forms for the suggested paradigms in order to *find support for the paradigms in a corpus*. In Section 2, we outline the directly related previous work. In Section 3, we describe the new method. In Section 4, we present the training and test data. In Section 5, we evaluate the model. In Section 6, we discuss the method and the test results in light of the existing literature and some similar methods.

2 Lexicon-based Entry Generator

To create entries for a morphological analyzer from previously unseen words, we need an entry generator. Ideally, we can use information that is already available in some existing morphological description to encode new entries in a similar fashion. Below, we briefly outline a general method for creating lexicon-based entry generators that was introduced by Lindén (2009). In his article, Lindén demonstrates that the method works well for English, Finnish and Swedish.

Assume that we have a finite-state transducer lexicon T which relates base forms, $b(w)$, to inflected words, w. Let w belong to the input language L_I and $b(w)$ to the output language L_O of the transducer lexicon T. Our goal is to create an entry generator for inflected words that are unknown to the lexicon, i.e. we wish to provide the most likely base forms $b(u)$ for an unknown input word $u \notin L_I$. In order to create an entry generator, we first define the left quotient and the weighted universal language with regard to a lexical transducer. For a general introduction to automata theory and weighted transducers, see e.g. Sakarovitch (2003).

If L_1 and L_2 are formal languages, the left quotient of L_1 with regard to L_2 is the language consisting of strings w such that xw is in L_1 for some string x in L_2. Formally, we write the left quotient as in Equation 1.

$$L_1 \setminus L_2 = \{ a \mid \exists x ((x \in L_2) \wedge (xa \in L_1)) \} \quad (1)$$

We can regard the left quotient as the set of postfixes that complete words from L_2, such that the resulting word is in L_1.

If L is a formal language with alphabet Σ, a universal language, U, is a language consisting of strings in Σ^*. The weighted universal language, W, is a language consisting of strings in Σ^* with weights $p(w)$ assigned to each string. For our purposes, we define the weight $p(w)$ to be proportional to the length of w. We define a weighted universal language as in Equation 2.

$$W = \{ w \mid \exists w (w \in \Sigma) \} \quad (2)$$

with weights $p(w) = C \mid w \mid$, where C is a constant.

A finite-state transducer lexicon, T, is a formal language relating the input language L_I to the output language L_O. The pair alphabet of T is the set of input and output symbol pairs related by T. An identity pair relates a symbol to itself.

We create an entry generator, G, for the lexicon T by constructing the weighted universal language W for identity pairs based on the alphabet of L_1 concatenating it with the left quotient of T with regard to the universal language U of the pair alphabet of T as shown in Equation 3.

$$G(T) = W\ T \setminus \mathrm{U} \quad (3)$$

Lindén (2009) proves that it is always possible to create an entry generator, $G(T) = W\ T \setminus \mathrm{U}$, from a weighted lexical transducer T.

The model is general and requires no information in addition to the lexicon from which the entry generator is derived. Therefore Lindén suggests that it be used as a baseline for other entry generator methods.

3 Corpus-based Paradigm Selection

To score the top paradigms suggested by an entry generator, we generate some of the key word forms of a paradigm and compare them against a corpus. A paradigm whose key word forms are well-attested, i.e. used many times, is more likely to be correct than a paradigm whose word forms

[1] http://www.lingsoft.fi/

only have a few documented cases. Rare forms may even be spelling errors. By scoring all the paradigms provided by the paradigm guesser according to the frequency of the word forms and then comparing the scores, we find the paradigm that is most likely to be correct.

We define a method for scoring possible paradigms of an unknown word. Let us define a set of paradigms of an unknown word $U_p = \{P_1, P_2, P_3, ...P_n\}$. Each paradigm P_n has a set that consists of the paradigm's key words, $W_n = \{w_1, w_2, w_3, ...w_m\}$. A distinct word form w_K may simultaneously belong to the key word sets of several paradigms.

Each distinct word form w_K has a number of occurrences $o_c(w_K)$ in the corpus. If a key word belongs to the key word sets of more than one paradigm, the key word does not differentiate well between those paradigms. Therefore each key word w_m only receives a score o_{w_m} equal to the number of occurrences $o_c(w_K)$ in the corpus divided by the number $o_p(w_m)$ of key words w_m matching w_K in the set of paradigms U_P. The score of a key word is defined in Equation 4.

$$o_{w_m} = \frac{o_c(w_K)}{o_p(w_m)}, \quad w_K = w_m \tag{4}$$

We add the scores, o_w, of the key words in a paradigm and divide the sum by the number, $|W_p|$ of key words in the paradigm. The score of a paradigm is defined in Equation 5:

$$Score_{P_n} = \frac{\sum O_w}{|W_p|}, w \in W_n \tag{5}$$

A key word form can have several variants, e.g. the genitive plural of Finnish nouns may have up to three different variants for each word in a paradigm. The variants all represent a single word form, i.e. genitive plural. We select the largest variant score to represent the score of the word form.

The method orders the suggestions from the entry generator. If the method does not differentiate between two suggestions, the order proposed by the generator prevails.

The method can be used with any data that reflects the occurrence of the paradigm key words. Although we refer to the source of word frequency data as a corpus, the method can be used with other data sources as well. As is described in section 5, we have successfully tested the method using both corpus material and page frequencies returned by a web search engine. In theory, the method should work with any data source that reflects the occurrence of words in language use.

4 Training and Test Data

To test our method for corpus-based paradigm selection of paradigms generated by a lexical entry generator, we used the entry generator for Finnish created by Lindén (2009) implemented with the *Helsinki Finite-State Technology* (HFST, 2008). In 4.1, we briefly describe the lexical resources used for the finite-state transducer lexicon which was subsequently converted into an entry generator.

Words unknown to the lexicon were drawn from a language-specific text collection. The correct entries for a sample of the unknown words were manually determined. In 4.2, we describe the text collections and the sample used as test data. In 4.3, we describe the evaluation method and characterize the baseline.

4.1 Lexical Data for a Finnish Finite-State Transducer Lexicon and Entry Generator

Lexical descriptions relate look-up words to other words and indicate the relation between them. A morphological finite-state transducer lexicon relates a word in dictionary form to all its inflected forms. For an introduction, see e.g. Koskenniemi (1983).

Our current Finnish morphological analyzer was created by Pirinen (2008) based on the Finnish word list *Kotimaisten kielten tutkimuskeskuksen nykysuomen sanalista* (2007), which contains 94 110 words in base form. Of these, approximately 43 000 are non-compound base forms classified with paradigm information. The word list consists of words in citation form annotated with paradigm and gradation pattern. There are 78 paradigms and 13 gradation patterns. For example, the entry for käsi (= 'hand') is 'käsi 27' referring to paradigm 27 without gradation, whereas the word pato (= 'dam') is given as 'pato 1F' indicating paradigm 1 with gradation pattern F. From this description a lexical transducer is compiled with a cascade of finite-state operations. For nominal paradigms, i.e. nouns and adjectives, inflection includes case inflection, possessive suffixes and clitics creating more than 2 000 word forms for each nominal. For the verbal inflection, all tenses, moods and personal forms are counted as inflections, as well as all infinitives and participles and their correspond-

ing nominal forms creating more than 10 000 forms for each verb. In addition, the Finnish lexical transducer also covers nominal compounding.

This finite-state transducer lexicon was converted into an entry generator using the procedure outlined in Section 2

4.2 Test Data

As test data, we use the *Finnish Text Collection,* which is an electronic document collection of the Finnish language. It consists of 180 million running text tokens. The corpus contains news texts from several current Finnish newspapers. It also contains extracts from a number of books containing prose text, including fiction, education and sciences. Gatherers are the Department of General Linguistics, University of Helsinki; The University of Joensuu; and CSC–Scientific Computing Ltd. The corpus is available through CSC [www.csc.fi].

We use the same test data as Lindén (2009), which is a set of previously unseen words in inflected form for which we wish to determine the inflectional paradigm. In order to extract word forms that represent relatively infrequent and previously unseen words, 5000 word and base form pairs had been drawn at random from the frequency rank 100 001-300 000. To get new words, only inflected forms that were not recognized by the lexical transducer were kept. However, from the test data, strings containing numbers, punctuation characters, or only upper case characters were also removed, as such strings require other forms of preprocessing as well in addition to some limited morphological analysis.

1. **ulkoasu** *1 noun* (appearance)
   ulkoasu ulkoasun ulkoasua ulkoasuun
   ulkoasut ulkoasujen ulkoasuja ulkoasuihin
2. **ulkoasu** *2 noun* (appearance)
   ulkoasu ulkoasun ulkoasua ulkoasuun ulkoasut
   ulkoasujen~ulkoasuitten~ulkoasuiden
   ulkoasuja~ulkoasuita ulkoasuihin
3. **ulkoasullata** *73 I verb* (to stuff sth from the outside)
   ulkoasullata ulkoasultaan ulkoasultasi
   ulkoasultaisi ulkoasullannee ulkoasullatkoon
   ulkoasullannut ulkoasullattiin
4. **ulkoasu** *21 noun* (appearance)
   ulkoasu ulkoasun ulkoasuta ulkoasuhun
   ulkoasut ulkoasuiden ulkoasuita ulkoasuihin

Picture 1. Word form *ulkoasultaan* (= by its appearance) and the combinations of **base form**, *paradigm information*, (English gloss added for readability of this picture only) and *key word forms* to be selected from.

Of the randomly selected strings, 1715 represented words not previously seen by the lexical transducer. For these strings, correct entries were created manually. Of these, only 48 strings had a verb form reading. The rest were noun or adjective readings. Only 43 had more than one possible reading.

A sample of test strings are: *ulkoasultaan* (by its appearance), *euromaan* (of the euroland), *työvoimapolitiikka* (labour market policy), *pariskunnasta* (from the couple), *vastalausemyrskyn* (of the protest storm), *ruuanlaiton* (of the cookery), *valtaannousun* (of the rise to power), *suurtapahtumaan* (for the major event), …

In Picture 1, we see an example of the word form *ulkoasultaan* and the suggested paradigms as they have been generated by the entry generator and expanded with key word forms in order for an evaluator to determine the correct paradigm for the morphological entry.

4.3 Evaluation Measures, Baselines and Significance Test

We report our test results using recall and average precision at maximum recall. *Recall* means all the inflected word forms in the test data for which an accurate base form suggestion is produced. *Average precision at maximum recall* is an indicator of the amount of noise that precedes the intended paradigm suggestions, where *n* incorrect suggestions before the *m* correct ones give a precision of *1/(n+m)*, i.e., no noise before a single intended base form per word form gives 100 % precision on average, and no correct suggestion at maximum recall gives 0 % precision. The *F-score* is the harmonic mean of the recall and the average precision.

The random baseline for Finnish is that the correct entry is one out of 78 paradigms with one out of 13 gradations, i.e. a random correct guess would on the average end up as guess number 507.

As suggested by Lindén (2009), we use the automatically derived entry generator from Section 4.1 as a baseline. Using his test data, the test results will be directly comparable to the baseline provided in Table 1 with recall 82 %, average precision 76 % and the F-score 79 %.

The significance of the difference between the baselines and the tested methods is tested with matched pairs. The Wilcoxon Matched-Pairs Signed-Ranks Test indicates whether the changes in the ranking differences are statistically significant. For large numbers the test is almost as sensitive as the Matched-Pairs Student t-test even if it does not assume a normal distribution of the ranking differences.

Rank	Freq	Percentage
#1	1140	66,5 %
#2	186	10,8 %
#3	64	3,7 %
#4	17	1,0 %
#5	4	0,2 %
#6	2	0,1 %
#7-∞	302	17,6 %
Total	1715	100,0 %

Table 1. Baseline for Finnish entry generator.

5 Evaluation

We test how well the entry selection procedure outlined in Section 3 is able to select the correct paradigm for an inflected word form using the test data described in Section 4.2. Word forms representing previously unseen words were used as test data in the experiment. The generated entries are intended for human post-processing, so the first correct entry suggestion should be among the top 6 candidates, otherwise the ranking is considered a failure. In 5.1, we test the paradigm selection procedure against a Finnish text corpus. In 5.2, we also test the paradigm selection procedure using page counts from the internet.

5.1 Corpus-based Paradigm Ranking

We evaluate the paradigm selection method on paradigms generated by the lexicon-based entry generator against the *Finnish Text Collection* described in Section 4.2.

Rank	Freq	Percentage
#1	1316	76,7 %
#2	110	6,4 %
#3	34	2,0 %
#4	25	1,5 %
#5	11	0,6 %
#6	9	0,5 %
#7-∞	210	12,2 %
Total	1715	100,0 %

Table 2. Ranks of all the first correct entries by the Finnish entry generator when ranking suggestions against the *Finnish Text Collection*.

The Finnish entry generator generated a correct entry among the top 6 candidates for 88 % of the test data as shown in Table 2, which corresponds to an average position of 1.9 for the first correct entry with 88 % recall and 83 % average precision, i.e. an 86 % F-score.

5.2 Page Count-based Paradigm Ranking

We also evaluate the paradigm selection method on paradigms generated by the lexicon-based entry generator against the Word-Wide Web using page counts for pages retrieved over a period of some weeks from Google for key words of the paradigms. We retrieved the data from pages which Google gave a Finnish language code. We used this as way to verify the method on an independent corpus.

Rank	Freq	Percentage
#1	1229	71,7 %
#2	115	6,7 %
#3	77	4,5 %
#4	28	1,6 %
#5	18	1,0 %
#6	11	0,6 %
#7-∞	231	13,5 %
Total	1715	100,0 %

Table 3. Ranks of all the first correct entries by the Finnish entry generator when ranking suggestions against the World-Wide Web.

The Finnish entry generator generated a correct entry among the top 6 candidates for 86 % of the test data as shown in Table 3, which corresponds to an average position of 2.1 for the first correct entry with 86 % recall and 79 % average precision, i.e. an 83 % F-score.

5.3 Significance

The selection of the paradigms from the morphological entry generator was statistically highly significantly better than the lexical baseline according to the Wilcoxon Matched-Pairs Signed-Ranks Test. The difference between the corpus and the internet might be statistically significant, but has no real practical implications. The improvement in the F-score of 4-8 percentage points from the baseline model in two separate test settings is significant in practice.

6 Discussion

In this section, we discuss the results and give a brief overview of some related work. In 6.1, we compare test results with previous efforts. In 6.2, we discuss future work.

6.1 Discussion of Results

The problem when dealing with relatively low-frequency words is that an approach to generate additional word forms for their paradigms may not contribute much. It might well be that the

word we are looking at is the only instance in the corpus. In that sense, turning to the internet for help seems like a good idea. It is interesting but not surprising to note that a relatively clean corpus still provides a slightly better basis for ranking word paradigms than the Internet. The most plausible explanation for this would be a larger amount of misspelled word forms which reduces the distinctions between paradigm suggestions, an effect that was observed during the evaluation.

Sometimes the misspelling was more common than the correctly spelled word. E.g., the sixth highest scoring word in our material was "seuraavä", with approx. 21 000 000 page counts, while its correctly spelled form, "seuraava", had almost 500 000 page counts less. This was in most cases corrected by a higher average frequency of the remaining word forms in the correct paradigm. Sometimes the incorrect paradigms happened to contain a homonym of some frequently occurring words, which raised the score of the paradigm above that of the correct paradigm candidate.

It is significant to note that our experiment demonstrates that the ranking can be performed using page counts instead of word counts with a sufficiently large corpus, which is by no means self-evident. Essentially page counts mean that we use the semantic context of a word. Many of the inflected forms will refer to the same pages, which also opens up avenues for future research. One could perhaps check how many pages contain the base form in addition to some inflected form of a paradigm in order to reduce the noise.

The fact that as a source of data, the corpus data fared slightly better than the internet may in our case also be attributable to the fact that Finnish word forms in the frequency range 100 000-300 000 may not be so rare after all due to the rich morphology and productive compounding mechanism of Finnish.

From a practical point of view, we are able to significantly reduce the workload of encoding lexical entries as most of the task can be accomplished automatically. However, a significant change is that assigning paradigms to words, which previously required an expert lexicographer, can now be accomplished by a native speaker making a choice, in practice, between the first two or at most three suggestions from the computer.[2]

[2] http://www.ling.helsinki.fi/cgi-bin/omor/omorfi-cgi-demo.py

6.2 Comparison with similar or related efforts

A related idea of expanding key word forms of paradigms to identify new words and their paradigms has been suggested by Hammarström & al (2006). However, their approach was to automatically deduce rules for which they could find as much support as was logically possible in order to make a safe inference. This leads to safely extracting words that already have a number of word forms in the corpus, i.e. mid- or high-frequency words, which for all practical purposes have already been encoded and are readily available in public domain morphological descriptions like the Ispell dictionaries (Kuenning, 2007) or more advanced descriptions like the Finnish dictionary *Kotimaisten kielten tutkimuskeskuksen nykysuomen sanalista* (2007). It should be noted that Hammarström & al (2006) came to the conclusion that it is recommendable that a linguist writes the extraction rules.

The approach suggested by Mikheev (1996, 1997) aims at solving the issue of unknown words in the context of part-of-speech taggers. However, in this context the problem is slightly easier as the guesser only needs to identify a likely part of speech and not the full inflectional paradigm of a word. He suggests an automatic way of extracting prefix and postfix patterns for guessing the part of speech. A related approach aiming at inducing paradigms for words and inflectional morphologies for 30 different languages is suggested by Wicentowski (2002).

Since there is a growing body of translated text even for less studied languages, there are interesting approaches using multi-lingual evidence for inducing morphologies, see e.g. Yarowski and Wicentowski (2000). This approach is particularly fruitful if we can use relations between closely related languages.

If we cannot find enough support for any particular paradigm of a word, e.g. if the word is too infrequent so that there are no other inflections, we need a way to make inferences based on related or similar strings. We need to make inferences based on the analogy with already known words as suggested e.g. by Goldsmith (2007) or Lindén (2008, 2009).

6.3 Future Work

The current approach only extracts inflectional information in the form of paradigms, even if the context of a new word also contributes other types of lexical information such as part of

speech, argument structure and other more advanced types of syntactic and semantic information.

The Internet as a source of data also provides context for a search word, some of it specific to this particular data source. Our current approach does not yet take into account the nature of this source of data, such as an increased occurrence of misspellings, colloquial word forms and mixed-language content. Also, as the Internet is an ever-changing medium, any linguistic data derived form it is subject to constant change. The effect of this change to the reliability of evaluation needs to be further investigated.

7 Conclusions

We have proposed and successfully tested a new method for selecting paradigms generated for inflected forms of new words using additional corpus information for key forms of the paradigms suggested by en entry generator. We tested the model on Finnish, which is a highly inflecting language with a considerable set of inflectional paradigms and stem change categories. Our model achieved 79-83 % precision and 86-88 % recall, i.e. an F-score of 83-86 %. The average position for the first correctly generated entry was 1.9-2.1. The method was highly statistically significantly better than a non-trivial baseline and the improvement is also significant in practice.

Acknowledgments
We are grateful to the Finnish Academy and to the Finnish Ministry of Education for the funding and to the members of the HFST team at the University of Helsinki for fruitful discussions.

References

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., Stolcke, A. 2007 Morph-based speech recognition and modeling of out-of-vocabulary words across languages. In *ACM Transactions on Speech and Language Processing*, 5(1) article 3.

Forsberg, M., Hammarström, H., and Ranta, A. 2006. Morphological Lexicon Extraction from Raw Text Data. *FinTAL 2006*, LNCS 4139, pages 488-499, 2006.

Goldsmith, J. A. 2007. Morphological Analogy: Only a Beginning, http://hum.uchicago.edu/ ~jagoldsm/ Papers/analogy.pdf

HFST–Helsinki Finite-State Technology. 2008. http://www.ling.helsinki.fi/kieliteknologia/ tutkimus/hfst/index.shtml

Koskenniemi, K.. 1983. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. PhD Thesis. Department of General Linguistics, University of Helsinki, Publication No. 11.

Kotimaisten kielten tutkimuskeskuksen nykysuomen sanalista, 2007. Reseach Institute for the Languages of Finland. http://kaino.kotus.fi/sanat/ nykysuomi/

Kuenning, G. 2007 Dictionaries for International Ispell, http://www.lasr.cs.ucla.edu/geoff/ispell-dictionaries.html

Kurimo, M., Creutz, M., Turunen, V. 2007. Overview of Morpho Challenge in CLEF 2007. In *Working Notes of the CLEF 2007 Workshop*, pages 19-21.

Lindén, K. 2008. A Probabilistic Model for Guessing Base Forms of New Words by Analogy**.** In *9th International Conference on Intelligent Text Processing and Computational Linguistics*, Haifa, Israel, LNCS 4919, pages 106-116.

Lindén, K. 2009. Guessers for Finite-State Transducer Lexicons**.** In *CICling-2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics*, March 1-7, 2009, Mexico City, Mexico.

Mikheev, A. 1997. Automatic Rule Induction for Unknown-Word Guessing. In *Computational Linguistics,*. 23(3), pages 405-423.

Mikheev, A. 1996. Unsupervised Learning of Word-Category Guessing Rules. In: *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, pages 327-334.

Oflazer, K., Nirenburg, S., McShane, M. 2001. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. In *Computational Linguistics*, 27(1), pages 59-85.

Pirinen, T. 2008. Open Source Morphology for Finnish using Finite-State Methods (in Finnish). Technical Report. Department of Linguistics, University of Helsinki.

Sakarovitch, J. 2003. *Éléments de théorie des automates.* Vuibert

Wicentowski, R. 2002. Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework. PhD Thesis, Baltimore, USA.

Yarowsky, D. and Wicentowski. R. 2000. Minimally Supervised Morphological Analysis by Multimodal Alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.*

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206