

# Guessers for Finite-State Transducer Lexicons

Krister Lindén

Department of General Linguistics, P.O. Box 9, FIN-00014 University of Helsinki  
Krister.Linden@Helsinki.fi

**Abstract.** Language software applications encounter new words, e.g., acronyms, technical terminology, names or compounds of such words. In order to add new words to a lexicon, we need to indicate their inflectional paradigm. We present a new generally applicable method for creating an entry generator, i.e. a paradigm guesser, for finite-state transducer lexicons. As a guesser tends to produce numerous suggestions, it is important that the correct suggestions be among the first few candidates. We prove some formal properties of the method and evaluate it on Finnish, English and Swedish full-scale transducer lexicons. We use the open-source *Helsinki Finite-State Technology* [1] to create finite-state transducer lexicons from existing lexical resources and automatically derive guessers for unknown words. The method has a recall of 82-87 % and a precision of 71-76 % for the three test languages. The model needs no external corpus and can therefore serve as a baseline.

## 1 Introduction

New words and new usages of old 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 some applications, hand-made guessers are used for covering the low-frequency vocabulary or the strings are simply added as such.

Mikheev [2] and [16] noted that words unknown to the lexicon present a substantial problem to part-of-speech tagging and he presented a very effective supervised method for inducing a guesser from a lexicon and an independent training corpus. Oflazer & al. [3] 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 [4] and Goldsmith [5]. 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 [6]. For a comparison of some recent successful segmentation methods, see the Morpho Challenge [7].

Although unsupervised methods have advantages for less-studied languages, for the well-established languages, we have access to fair amounts of lexical training ma-

terial in the form of analyzes 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 shown by Lindén [8] 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. In addition, there are a host of large but shallow hand-made morphological descriptions available, e.g., the Ispell collection of dictionaries [9] for spell-checking purposes, and many well-documented morphological analyzers are commercially available, e.g. [10].

In this paper, we propose a new method that takes an existing finite-state transducer lexicon and creates a guesser using only generally applicable formal properties of weighted transducers. The method is implemented using the open-source *Helsinki Finite-State Technology* [1]. In Section 2, we describe the methodology and present some formal properties. In Section 3, we present the training and test data. In Section 4, we evaluate the model on Finnish, English and Swedish transducer lexicons. In Section 5, we discuss the method and the test results in light of previous literature on guessers.

## 2 Methodology

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  $L_1$  and  $b(w)$  to the output language  $L_0$  of the transducer lexicon. Our goal is to create a guesser 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  $u \notin L_1$ .

In 2.1, we describe the theoretical foundation of the guesser model and, in 2.2, we prove some of the fundamental properties of the guesser model.

### 2.1 Guesser Model

In order to create a guesser, 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 [23].

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  $L_2$ . In symbols, we write the left quotient as:

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

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  $\Sigma$ , a *universal language*,  $U$ , is a language consisting of strings in  $\Sigma^*$ . The *weighted universal language*,  $W$ , is a language consisting of strings in  $\Sigma^*$  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:

$$W = \{ w \mid \exists w (w \in \Sigma^*) \} \text{ with weights } p(w) = C \cdot |w|, \quad (2)$$

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 a guesser,  $G$ , for the lexicon  $T$  by constructing the weighted universal language  $W$  for identity pairs based on the alphabet of  $L_I$  concatenating it with the left quotient of  $T$  for the universal language  $U$  of the pair alphabet of  $T$ :

$$G(T) = WT \setminus U \quad (3)$$

## 2.2 Properties of the Guesser Model

**Lemma 1.** For the lexicon,  $T$ , a guesser,  $G(T)$ , composed with an unknown word,  $u$ , generates the entry guesses  $b(u) = y b(w)$ , where  $b(w)$  is a postfix of  $L_O$  and  $w$  is a postfix of  $L_I$ .

*Proof.* Assume that we have an unknown word  $u \in \Sigma^*$ , where  $\Sigma$  is the input alphabet of  $T$ . We decompose  $u$  into  $yw$ , such that  $y \in \Sigma^*$  and  $w \in \{s \mid \exists p ((p \in L_I) \wedge (ps \in \Sigma^*))\}$ . We then have  $u \circ G(T) = (yw) \circ (WT \setminus U) = (y \circ W)(w \circ T \setminus U) = y b(w)$ .  $\square$

**Lemma 2.** For the weight-free transducer,  $T$ , the entry guesses with the minimal weight,  $b_{min}(u)$ , for an unknown word,  $u$ , composed with the guesser,  $G(T)$ , is generated by the set of longest matching postfixes of  $u$  and the input language of  $T$ .

*Proof.* Assume that we have an unknown word  $u \in \Sigma^*$ , where  $\Sigma$  is the input alphabet of  $T$ . The  $b(u)$  with minimal weight is  $b_{min}(u) = \arg \min_{p(v)} \{ v \mid v = u \circ G(T) \} = \arg \min_{p(v)} \{ v \mid v = y b(w) \}$ . The weight of  $y b(w)$  is proportional to the length of  $y$ , i.e.  $|y| \cdot C$ .  $\square$

**Lemma 3.** If  $T$  is a weighted transducer, the guesses with minimal weight,  $b_{min}(u)$ , are the longest matching postfixes of  $u$  with minimal weight by  $T$  provided that the weight  $C$  of the symbols in the universal language  $W$  is greater than the weight of any symbol pair related by  $T$ .

*Proof.* Assume that we have an unknown word  $u \in \Sigma^*$ , where  $\Sigma$  is the input alphabet of  $T$ . We decompose  $u$  into  $y_1 w_1$  and  $y_2 w_2$ , such that  $y_1, y_2 \in \Sigma^*$  and  $w_1, w_2 \in \{s \mid \exists p ((p \in L_I) \wedge (ps \in \Sigma^*))\}$  and  $|y_1| = |y_2| - 1$ . As  $|w_1| = |w_2| + 1$ , we have

$p(b(w_1)) \geq p(b(w_2)) + C$  and consequently  $p(y_1 b(w_1)) \geq p(y_2 b(w_2))$ .

□

**Theorem.** To create a longest matching postfix guesser,  $G(T) = W T \setminus U$ , from the weighted lexical transducer,  $T$ , we take the maximum transition weight,  $\omega$ , of  $T$  and assign the prefix transition weight  $C$  to  $\omega + \delta$ .

*Proof.* The result follows directly from Lemma 3.

□

For prefixing languages, we can create a guesser using the right quotient and the universal postfix. For circumfixing languages, we can concatenate the prefixing and postfixing guessers to create a circumfixing guesser.

Generally, one can characterize our weighted finite-state entry generator as inducing an ordering over the possible entries for a new and previously unseen inflected form preferring entries that have inflected forms and parts of the stem in common with previously seen entries. As a corollary, entries for already seen words will be generated first. If the forms of the lexical transducer,  $T$ , are weighted according to the frequency of the paradigms in the lexicon, the most frequent paradigms are generated first if there are several paradigm candidates for the same affix.

### 3 Data Sets

To test the entry generator for finite-state transducer lexicons, we created transducer lexicons from existing lexical resources for three different languages: Finnish, Swedish and English using the *Helsinki Finite-State Technology* [1]. We drew words unknown to these lexicons from three language-specific text collections and manually determined their correct entries. In 3.1, we describe the lexical resources and outline the procedure for creating the finite-state transducer lexicon. In 3.2, we describe the test data and, in 3.3, we describe the evaluation method and characterize the baselines.

#### 3.1 Lexical Data for Finite-State Transducer Lexicons

The lexical descriptions relate base forms to inflected word forms. This can be done either through each base form classified with a paradigm and a list of paradigms with model words, or it can be done as a full-form lexical description with all the inflected forms of each base form. The final lexicon and is implemented with finite-state transducer technology. Regardless of the initial form of the lexical description, the finite-state transducer lexicon maps a word in dictionary form to all of its inflected forms. For an introduction, see e.g. Koskenniemi [11]. Essentially this means that composing the transducer lexicon with an inflected word form will create a new transducer containing all the possible base forms and the morphological analyses of how the inflected word form is related to the base form.

A weighted finite-state transducer lexicon can contain weights in many different ways. A fruitful set of weights would be to estimate the relative frequency of the word

forms and encode them as a priori probabilities or weights in the lexicon. This requires a disambiguated corpus. Above we only have lexical descriptions and, assuming that there are or we have created inflectional paradigms, we can estimate the relative frequency of the paradigms. It has also been demonstrated by Karlsson [12] that it is preferable to have as few parts as possible in a multipart compound analysis. For lack of better estimates, the weighted finite-state transducer lexicon lists the analyses primarily according to the number of analyzed compound parts and secondarily in paradigm frequency order.

Most languages have ready-made inflectional paradigms with the lexical description. From this a finite-state transducer lexicon can be manually compiled. However, for languages which typically have few inflected forms for each base form, it is feasible to have a full-form description of all the lexical entries. If we only have a full-form lexical description, we need to extract paradigms, in order to be able to generate lexical entries for new words.

**Finnish.** In order to create the Finnish dictionary, we used the Finnish word list *Nyky-suomen sanalista* [13], 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 with 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 [22]. For nominal paradigms, 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 corresponding nominal forms creating more than 10 000 forms for each verb. In addition, the Finnish lexical transducer also covers nominal compounding.

**English.** For English we use *FreeLing 2.1* [14]. The FreeLing English lexical resource was automatically extracted from WSJ, with manual post-editing and completion. It contains about 55 000 forms corresponding to some 40 000 different combinations of lemma and part-of-speech. For each part-of-speech, English only has a small set of forms for phonological or semantic reasons, but most often due to the fact that the form did not occur in the Brown corpus.

We extract paradigms from the full-form lexical description for English in the following manner: we automatically align the characters of the base form and the inflected forms and determine the longest common prefix for the base form and all the inflected forms. The remaining set of endings, possibly with some characters from the stem, is considered a paradigm. Since the words may have individual patterns with left out forms, the automatically extracted set of paradigms becomes relatively large. We get 489 paradigms for English out of which 151 occur more than once.

**Swedish.** For Swedish we use the open source full-form dictionary *Den stora svenska ordlistan* [15]. For each base form, the part of speech is given. For each part-of-speech, there is a given set of inflected forms, e.g. for nouns there are always eight

forms, i.e. all combinations of singular and plural, nominative and genitive, definite and indefinite forms. For any word, there may be an empty slot, if the form is considered non-existent for some reason, e.g. phonologically or semantically. In addition, each word may have an indication of whether it can take part in compounding which is prolific in Swedish.

We use the same procedure for creating paradigms for Swedish as we used for English. We get 1333 paradigms out of which 544 occur more than once with the rest in a Zipfian distribution.

### 3.2 Test Data

A set of previously unseen words in inflected form serve as a test set for which we wish to determine their inflectional paradigm. In order to extract word forms that represent relatively infrequent and previously unseen words we used various text collections for Finnish, Swedish and English. We drew 5000 word and base form pairs at random from the frequency rank 100 001-300 000 as test material for each language. Since we are interested in new words, we only counted inflected forms that were not recognized by the lexical transducers we had created. In addition, we removed strings containing numbers, punctuation characters or only upper case from the test data.

**Finnish.** For Finnish, we used the *Finnish Text Collection*, which is an electronic document collection of the Finnish language. It consisted 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](http://www.csc.fi)].

Of the 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, kilpailulainsäädännön, epätasa-arvoa, euromaan, työvoimapolitiikka, pariskunnasta, vastalausemyrskyn, kolmeentoista, haudatut, liioitellun, ruuanlaiton, valtaannousun, suurtahtumaan, ostamiaan, ...*

**English.** For English, we used part of *The Project Gutenberg* text collection, which consists of thousands of books. For this experiment we used the English texts released in the year 2000 [<http://www.gutenberg.org/>]. The tokens consisted of 266 000 forms of 175 000 base forms.

Of the selected strings, 3100 represented words not previously seen by the lexical transducer. For these strings, correct entries were created manually for the first 25 %, i.e. 775 new entries. Of these, 60 strings had verb form readings, 610 noun readings and 161 adjective readings, and 14 adverb readings. Only 79 strings had more than one reading.

A sample of test strings are: *florin, disfranchised, chimney-pieces, Beechwood, warbled, sureness, sitting-rooms, marmoset, landscape-painter, half-burnt, Burlington, ...*

**Swedish.** For Swedish, we used the *Finnish-Swedish Text Collection*, which is an electronic document collection of the Swedish language of the Swedish speaking minority in Finland. It consisted of 35 million tokens. The tokens were 765 000 inflected forms of 445 000 base forms. The corpus contains news texts from several current Finnish-Swedish newspapers. It also contains extracts from a number of books containing fiction prose text. Gatherers are The Department of General Linguistics, University of Helsinki; CSC–Scientific Computing Ltd. The corpus is available through CSC [www.csc.fi].

Of the selected strings, 1756 represented words not previously seen by the lexical transducer. For these strings, correct entries were created manually for first 25 %, i.e. 439 new entries. Of these, 37 strings had a verb form reading, 387 noun readings, 47 adjective readings. Only 48 strings had more than one reading.

A sample of the test strings are: *finrummet, chansons, översvämmande, Valören, tonsiller, Stollans, sjöfartspolitiska, relikten. oskött. Dylikt, antidopingkommitté, ...*

### 3.4 Evaluation Measures and Baseline

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 base form 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 precision and the recall.

The random baseline for Finnish is that the correct entry is one out of the 78 paradigms with one out of 13 gradations, i.e. a random correct guess would on the average end up in as guess number 507. For English, an average random guess ends up in position 245 and, for Swedish, in position 667.

## 4 Experiments

We test how well the guesser outlined in Section 2 is able to predict the paradigm for an inflected word form using the test data mentioned in Section 3. Of the randomly chosen strings from the test data range, 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. All the guessers were statistically highly significantly better than their random baseline.

#### 4.1 Finnish Guesser

The Finnish Guesser generated a correct entry among the top 6 candidates for 82 % of the test data as shown in Table 1, which corresponds to an average position of 2.3 for the first correct entry with 82 % recall and 76 % average precision.

**Table 1.** Ranks of all the first correct entries by the Finnish guesser.

<i>Rank</i>	<i>Freq</i>	<i>Percentage</i>
#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 %
<b>Total</b>	1715	100,0 %

**Table 2.** Ranks of all the first correct entries by the English guesser.

<i>Rank</i>	<i>Freq</i>	<i>Percentage</i>
#1	477	61,5 %
#2	81	10,5 %
#3	56	7,2 %
#4	17	2,2 %
#5	14	1,8 %
#6	15	1,9 %
#7-∞	115	14,8 %
<b>Total</b>	775	100,0 %

**Table 3.** Ranks of all the first correct entries by the Swedish guesser.

<i>Rank</i>	<i>Freq</i>	<i>Percentage</i>
#1	243	55,4 %
#2	84	19,1 %
#3	40	9,1 %
#4	10	2,3 %
#5	5	1,1 %
#6	1	0,2 %
#N-∞	56	12,8 %
<b>Total</b>	439	100,0 %

## 4.2 English Guesser

The English Guesser generated a correct entry among the top 6 candidates for 83 % of the test data as shown in Table 2, which corresponds to an average position of 2.4 for the first correct entry with 83 % recall and 72 % average precision.

## 4.3 Swedish Guesser

The Swedish Guesser generated a correct entry among the top 6 candidates for 87 % of the test data as shown in Table 3, which corresponds to an average position of 2.3 for the first correct entry with 87 % recall and 71 % average precision.

# 5 Discussion

In this section, we give a brief overview of previous and related work on guessers. In 5.1, we compare test results with previous efforts. In 5.2, we give some notes on the implementation of the methods. In 5.3, we discuss future work.

## 5.1 Comparison with Results from Similar Efforts

Test results on identical data are not available, but similar efforts have been made and some insights can be gleaned from a comparison between them.

Stroppa and Yvon [17] present experimental results obtained on a morphological analysis task guessing base form and morphological features for an inflected form in English with the following recall and precision: nouns 75 % and 95 %; verbs 95 % and 97 %; adjectives 28 % and 88 %, respectively. It is interesting to note that verb forms are the easiest to get right, whereas it is much trickier to guess the base forms and syntactic features of nouns and adjectives. The explanation is probably that the base forms of nouns and adjectives are much more varied, and that they partly overlap with the inflected forms.

Wicentowski [18] presents the WordFrame model, a noise-robust supervised algorithm capable of inducing morphological analyses for languages which exhibit prefixation, suffixation, and internal vowel shifts. In combination with a naive approach to suffix-based morphology, this algorithm is shown to be remarkably effective across a broad range of languages, including those exhibiting infixation and partial reduplication. Results are presented for over 30 languages with a median accuracy of 97.5 % on test sets including both regular and irregular *verbal* inflections. The excellent accuracy is partly explained by the fact that he uses a dictionary to filter the suggested base forms. His intention is to learn irregular forms which are dominant among verbal inflections, but the good results should be seen in light of the results from Yvon and Stroppa [17], where a substantial challenge seems to be in modeling the behavior of nouns and adjectives. They are also the most frequent categories among new words.

Claveau and L’Homme [19] label morphologically related words with their semantic relations using morphological prefix and postfix analogies learned from a sample of pre-labeled words with a recall of 72 % and precision of 65 % on separate test data.

Baldwin [20] acquires affix and prefix transformations achieving 0.6 F-score for English using Timbl [21] as the classifier. However, the classification was for syntactic features not for inflectional paradigm.

We recall that our model is developed for guessing the paradigms of unknown and previously unseen inflected words, i.e. their base forms cannot be tested against a lexicon. In light of the results from comparable reports from other languages, our results automatically derived guessers are very good, because the data shows that the final guessers have 68-73 % precision and 82-87 % recall, i.e. an F-score of 78-79 %, on all three languages with different morphological complexity. It is interesting that our model is slightly more precise for Finnish, which is morphologically more complex than Swedish and English, whereas the recall is lower for Finnish. The explanation may be that inflected forms of Finnish are better indicators of the paradigm to which they belong, if the ending is recognized. In English, word endings may occur both in inflected and in base forms, e.g. ‘sleeping’ should be regarded as an adjective in base form in ‘a sleeping beauty’, but as an inflected form of the verb ‘sleep’ in ‘is sleeping’.

A quick look at the words which fail for English reveals that among them are e.g. *preacheth*, *Surmountheth*, *corrupteth*, which could not have received a correct guess as out-dated verb forms were not available as analogical models. Other words with missing correct analogues are *webbed*, which gets the base form *webb* whereas we expect *web*, even if the word is otherwise correctly identified as a verb. Similar problems seem to afflict words ending in low frequency characters in combination with the fact that we require the correct answer to have a specific base form and a paradigm which indicates all the correct inflected forms. E.g., we require that words like *plowman* are correctly identified as having the plural *plowmen*. It is not enough just to identify it is as some noun. The same goes for other words with irregular forms or exceptional paradigms. This also demonstrates that for part-of-speech tagging, the guessing task is easier as the tagging does not require guessing all the correct forms and only the correct forms of an out-of-vocabulary word.

**Table 4.** Test results for Finnish, English and Swedish guessers.

<i>Language</i>	<i>Recall</i>	<i>Precision</i>	<i>F-score</i>
Finnish	0,82	0,76	0,79
English	0,83	0,72	0,78
Swedish	0,87	0,71	0,78

## 5.2 Implementation note

The models were implemented with a cascade of weighted finite-state transducers. For conveniently creating morphological analyzers and guessers, HFST—the Helsinki Finite-State Technology [1] is available as an Open Source toolkit. Running the

guesser in forward mode may be relatively slow, whereas running the guesser in reverse is almost deterministic for the n-best results and therefore very efficient.

### 5.3 Future work

The suggested model is completely general and requires no additional data except the morphological analyzer in finite-state transducer format. It would be interesting to see, whether this general model can benefit from a purely probabilistic model conditioned on analogical transformations, e.g. the one suggested by Lindén [8], or some more contextually oriented model taking the surrounding words into account.

## 6. Conclusion

A substantial amount of languages have been implemented as lexical transducers with the Koskenniemi two-level model or similar formalisms, which means that there is a wealth of lexical transducers available. As the entry generator model we suggest is general and requires only a lexical transducer and no additional information from external corpora, it can serve as the baseline for entry generators on a number of languages. Compared with guessers for part-of-speech tagging, the entry guessing task is more difficult as entry guessing requires all the correct forms and only the correct forms of an out-of-vocabulary word to be identified. We have tested our entry guesser on inflected forms of new words in three languages from different language families demonstrating that the model has a recall of 82-87 % and a precision of 71-76 % for the three test languages. This corresponds to having the first correct entry on the average in position 2.3-2.4.

**Acknowledgements.** We are grateful to Tommi Pirinen and Anssi Yli-Jyrä for fruitful discussions and to the Finnish Academy for funding the research.

## References

1. HFST–Helsinki Finite-State Technology, <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/index.shtml>
2. Mikheev, A.: Unsupervised Learning of Word-Category Guessing Rules. In: Proc. of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96), pp 327-334 (1996)
3. Oflazer, K., Nirenburg, S., McShane, M.: Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. In: *Comp. Ling.*, vol. 27, no. 1, pp 59-85 (2001)
4. Wicentowski, R.: Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework. PhD Thesis, Baltimore, USA (2002)
5. Goldsmith, J. A.: Morphological Analogy: Only a Beginning, <http://hum.uchicago.edu/~jagoldsm/Papers/analogy.pdf> (2007)
6. Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., Stolcke, A.: Morph-based speech recognition and modeling of

- out-of-vocabulary words across languages. In: ACM Trans. on Speech and Lang. Proc., vol. 5, no. 1, art. 3 (2007)
7. Kurimo, M., Creutz, M., Turunen, V.: Overview of Morpho Challenge in CLEF 2007. In: Working Notes of the CLEF 2007 Workshop, pp. 19-21 (2007)
  8. Lindén, K.: 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, vol. 4919, pp. 106-116. Springer (2008)
  9. Kuenning, G.: Dictionaries for International Ispell, <http://www.lasr.cs.ucla.edu/geoff/ispell-dictionaries.html> (2007)
  10. Lingsoft, Inc.: *Demos*, [http://www.lingsoft.fi/?doc\\_id=107&lang=en](http://www.lingsoft.fi/?doc_id=107&lang=en)
  11. Koskeniemi, K.: Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics, University of Helsinki, Publication No. 11. (1983)
  12. Karlsson, F.: SWETWOL: A Comprehensive Morphological Analyser for Swedish. pp. 1–45, Nordic Journal of Linguistics, vol. 15, no. 1, Scandinavian University Press (1992)
  13. Nykysuomen sanalista, <http://kaino.kotus.fi/sanat/nykysuomi/>
  14. FreeLing 2.1—An Open Source Suite of Language Analyzers, <http://garraf.epsevg.upc.es/freeling/>
  15. Westerberg, T.: *Den stora svenska ordlistan*, <http://www.dssso.se/> (2008)
  16. Mikheev, A.: Automatic Rule Induction for Unknown-Word Guessing. In: Comp. Ling., vol. 23, no. 3, pp 405-423 (1997)
  17. Stroppa, N., Yvon, F.: An Analogical Learner for Morphological Analysis. In: Proc. of the 9th Conference on Computational Natural Language Learning (CoNLL), pp 120–127 (2005)
  18. Wicentowski, R.: Multilingual Noise-Robust Supervised Morphological Analysis using the WordFrame Model. In: Proc. of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology, ACL, pp 70-77 (2004)
  19. Claveau, V., L'Homme, M.C.: Structuring Terminology using Analogy-Based Machine Learning. In: Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, pp 17-18 (2005)
  20. Baldwin, T.: Bootstrapping Deep Lexical Resources: Resources for Courses. In: Proc. of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, ACL, pp 67-76 (2005)
  21. Daelemans, W., Zavrel, J., Sloot, K., Bosch, A: TiMBL: Tilburg Memory-Based Learner, version 6.0, Reference Guide', *Technical Report—ILK07-03*, Department of Communication and Information Sciences, Tilburg University (2003)
  22. Pirinen, T.: Open Source Morphology for Finnish using Finite-State Methods (in Finnish). Technical Report. Department of Linguistics, University of Helsinki. (2008)
  23. Sakarovitch, J. : Éléments de théorie des automates. Vuibert (2003)