A Probabilistic Model for Guessing Base Forms of New Words by Analogy

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, loan words, names or compounds of such words. Looking at English, one might assume that they appear in base form, i.e., the lexical look-up form. However, in more highly inflecting languages like Finnish or Swahili only 40-50 % of new words appear in base form. In order to index documents or discover translations for these languages, it would be useful to reduce new words to their base forms as well. We often have access to analyzes for more frequent words which shape our intuition for how new words will inflect. We formalize this into a probabilistic model for lemmatization of new words using analogy, i.e., guessing base forms, and test the model on English, Finnish, Swedish and Swahili demonstrating that we get a recall of 89-99 % with an average precision of 76-94 % depending on language and the amount of training material.

1 Introduction

New words and new usages of old words are constantly finding their way into daily language use. This is particularly prominent in quickly developing domains such as biomedicine and technology. Humans deal with new words based on previous experience: we treat them by analogy to known words. The new words are typically acronyms, technical terminology, loan words, names or compounds containing such words. They are likely to be unknown by most hand-made morphological analyzers. In some applications, hand-made guessers are used for covering this low-frequency vocabulary.

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 [8] and [2]. The problem is alleviated by the fact that there often are dictionaries available with common base forms or word roots for the most frequent words. If the inflectional patterns can be learned approximately from a corpus, the most common base forms can be checked against a dictionary in order to boost the performance of the methods. However, when we approach the other end of the spectrum, we have very rare words for which there are no ready base forms available in dictionaries and for heavily inflecting languages only 40-50 % of the words appear in base form in a

corpus. When new words appear for the first time, we also do not have access to several forms of the same word in order to draw on paradigmatic information.

If we do not need a full analysis, but only wish to segment the words into morph-like units, we can use a segmentation method like Morfessor [1]. For a comparison of some recent successful segmentation methods, see the Morpho Challenge [4].

Unsupervised methods have advantages for less-studied languages, but for the well-established languages, we have access to fair amounts of training material in the form of analyzes for more frequent words. There are a host of large but shallow hand-made morphological descriptions available, e.g., the Ispell collection of dictionaries [3] for spell-checking purposes, and many well-documented morphological analyzers are commercially available.

One can also argue that humans do not learn words by only observing masses of inflected forms. We are raised in a world, where we refer to similar objects and events using similar sound patterns. Context-based clustering methods have been proposed for this, but in lieu of more advanced methods for indicating words with identical or similar referents, we will use base forms for this purpose. We propose a new method for automatically learning a lemmatizer for previously unseen words, i.e., a base form guesser. This essentially places us in a supervised framework, but the novelty of the method is that we assume no knowledge of the structure of the morphology, i.e., the words could inflect word-initially or word-finally. In Section 2, we describe the probabilistic methodology. In Section 3, we present the training and test data for four morphologically distinct languages: English, Finnish, Swedish and Swahili. In Section 4, we test the model and show that the results are statistically very highly significant for all four languages. In Section 5, we discuss the method and the test results and give a note on the implementation.

2 Methodology

Assuming that we have set of word and base form pairs and another set of new previously unseen words for which we wish to determine their base forms by analogy with the known words, we first describe the probabilistic framework for our analogical model in Section 2.1. We then describe the probabilistic model for morphology in Section 2.2.

2.1 Probabilistic Framework for Analogy

Assume that we have a set of words, $w \in W$, from a text corpus for which we have determined the base forms, $b(w) \in B \subset W$, i.e. the lexicon look-up form. In addition, we have another set of words, $o \notin W$, for which we would like to determine their most likely

base form, $b(o) \notin B$. For this purpose, we use the analogy that w is to o as $b(w)$ is to $b(o)$. This relationship is illustrated in Figure 1.

$$\begin{array}{ccc} w & : & o \\ \updownarrow & & \updownarrow \\ b(w) & : & b(o) \end{array} \qquad\qquad \begin{array}{ccc} \textit{kokeella} & : & \textit{aikeella} \\ \updownarrow & & \updownarrow \\ \textit{koe} & : & ? \end{array}$$

Fig. 1. The analogy w is to o as $b(w)$ is to $b(o)$ illustrated by the Finnish words *kokeella ('with the test')* and *aikeella ('with the intention')*.

We use the analogical relation for deriving transformations $w \to b(w)$ from the differences between the known word and base forms. The transformations can then be applied to a new word o in order to generate a base form that should be similar to an existing base form $b(w)$. Several transformations may apply to any particular o and we wish to determine the most likely $b(o)$ in light of the evidence, i.e. we wish to find the $b(o)$, which maximizes the probability $P\left(o, w \to b(w), b(w), b(o)\right)$ for the new word o. By applying the chain rule to $P\left(o, w \to b(w), b(w), b(o)\right)$, we get (1),

$$\underset{b(o)}{\operatorname{argmax}} \sum_{b(w),\, w \to b(w)} \begin{pmatrix} P(o) \times \\ P\left(w \to b(w) \mid o\right) \times \\ P\left(b(w) \mid w \to b(w), o\right) \times \\ P\left(b(o) \mid b(w), w \to b(w), o\right) \end{pmatrix}. \qquad \textbf{(1)}$$

As the probability of $b(w)$ and o is independent of the other terms and the probability of o is constant, we get (2),

$$\underset{b(o)}{\operatorname{argmax}} \sum_{b(w),\, w \to b(w)} \begin{pmatrix} P\left(w \to b(w) \mid o\right) \times \\ P\left(b(w)\right) \times \\ P\left(b(o) \mid b(w), w \to b(w), o\right) \end{pmatrix}. \qquad \textbf{(2)}$$

We can also assume that the probability for a candidate base form $b(o)$ to be similar to existing base forms $b(w)$, is independent of the particular transformation $w \to b(w)$ that produced the candidate as well as of the source o of the candidate. In addition, we assume no knowledge of the distribution of the analog base forms $b(w)$, i.e. we assume an even distribution for maximum entropy. This further simplifies the expression to (3),

$$\underset{b(o)}{\operatorname{argmax}} \sum_{b(w),w\to b(w)} P\left(w \to b(w) \mid o\right) P\left(b(o) \mid b(w)\right) . \quad \textbf{(3)}$$

Finally, we make the Viterbi approximation and assume that the probabilities of the most likely transformations and the most similar base forms are good representatives of the sums of the probabilities over all transformations and base forms giving rise to the same candidate, which gives us the equation (4),

$$\underset{b(o)}{\operatorname{argmax}} P\left(w \to b(w) \mid o\right) P\left(b(o) \mid b(w)\right) . \quad \textbf{(4)}$$

We have now arrived at an expression that models the analogy between a word and its candidate base form in light of existing words and their base forms. The next step is to model the morphology of the words.

2.1 Probabilistic Framework for Morphology

When we have a model for calculating the analogy, we also need a model for the words and the morphological transformations that we can learn from the exemplars in a training corpus. We decompose the word o into consecutive substrings α, μ, ω and the candidate base form $b(o)$ into corresponding consecutive substrings β, ν, ξ, such that $\alpha \to \beta$ is a prefix transformation and $\omega \to \xi$ is a suffix transformation, whose likelihoods have been estimated from a set of pairs of words w and base forms $b(w)$ in a training corpus. (Note that the terms prefix, stem and suffix mean any string beginning, middle, or ending, respectively, and is not limited to the linguistically motivated terms prefix, stem and suffix morphemes.) Since we deal with new roots, the stem transformation $\mu \to \nu$ cannot be estimated from the corpus and needs a separate model, which we return to below. We assume that the prefix, stem and suffix, transformations can be applied independently. For the first part $P\left(w \to b(w) \mid o\right)$ of the analogy model (4), we get (5),

$$P\left(\alpha \to \beta \mid \alpha\mu\omega\right) P\left(\mu \to \nu \mid \alpha\mu\omega\right) P\left(\omega \to \xi \mid \alpha\mu\omega\right) . \quad \textbf{(5)}$$

Assume that we have a training corpus where the characters of the word and base form pairs are aligned. Estimating the probability of the prefix $P\left(\alpha \to \beta \mid \alpha\mu\omega\right)$ and suffix $P\left(\omega \to \xi \mid \alpha\mu\omega\right)$ transformations based on the aligned training data is straight forward. The conditional probability (6) of the prefix transformation is estimated directly from the counts $C\left(\alpha, \beta\right)$ of how often the prefixes α and β correspond in the aligned

word and base form data compared to the total count $C(\alpha)$ of the prefix α in the word forms,

$$P(\alpha \rightarrow \beta \mid \alpha\mu\omega) = \frac{C(\alpha,\beta)}{C(\alpha)} \ . \tag{6}$$

The conditional probability of the suffix transformations (7) for ω and ξ are estimated in the same way:

$$P(\omega \rightarrow \xi \mid \alpha\mu\omega) = \frac{C(\omega,\xi)}{C(\omega)} \ . \tag{7}$$

In (5), μ is likely to be a previously unseen stem, as we aim at modeling the inflections of new words. This means that we cannot really estimate its likelihood nor its transformations from a corpus. We therefore roughly model the new stem by a flat distribution (8) for the characters of the alphabet $\mu_i \in \mathrm{A}$ assuming that each character independently transforms only into itself, $\mu_i = \nu_i$. This stem model essentially assigns higher probability to shorter fragments of unknown stems. As a side-effect, we favor transformations for longer prefixes and suffixes. The size of the alphabet is $|\mathrm{A}|$ and the length of the stem μ is m.

$$P(\mu \rightarrow \nu \mid \alpha\mu\omega) = (1/|\mathrm{A}|)^m \ . \tag{8}$$

In order to model the likelihood of a new base form on its similarity to previously seen base forms, we compare the beginning and the end of a new base form to the base forms we have in the training material. Here β is a prefix and ξ is a suffix of a known base form and ν is a new stem. For the second part $P(b(o) \mid b(w))$ of the analogy model (4), we get (9),

$$P(\beta \mid \beta\nu\xi)\, P(\nu \mid \beta\nu\xi)\, P(\xi \mid \beta\nu\xi) \ . \tag{9}$$

Previously seen prefixes and suffixes of base forms are modeled with the conditional probability 1 and unseen prefixes and suffixes get the conditional probability 0. We again model the new word stem fragment by a flat distribution (10) for the characters of the alphabet $\nu_i \in \mathrm{A}$ assuming independence for the characters. The size of the alphabet is $|\mathrm{A}|$ and the length of the stem ν is n.

$$P(\nu \mid \beta\nu\xi) = (1/|\mathrm{A}|)^n \ . \qquad \textbf{(10)}$$

We now have all the components for a simple probabilistic model of inflectional morphology of unknown words, whose affix transformation parameters can be estimated from a character aligned training corpus of word and base form pairs. For details on how to align characters using a generalized edit distance alignment model, see e.g. [5].

3 Data Sets

In order to test our model in a language-independent setting, we selected four languages with different characteristics: *English*–a Germanic isolating language; *Swedish*–an agglutinating Germanic language; *Finnish*–a suffixing highly-agglutinating Fenno-Ugric language; *Swahili*–a prefixing language with a fair amount of suffixes as well. In Section 3.1, we present the corpora, from which we draw the training material as shown in Section 3.2 and test data as shown in Section 3.3. We present the baseline, measures and significance test in Section 3.4.

3.1 Corpus Data

We used publicly available text collections for the four languages: English, Finnish, Swedish and Swahili. An overview of the corpus sizes are displayed in Table 1.

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/dirs/GUTINDEX.00]. The texts were morphologically analyzed into 26 million running text tokens and disambiguated by the Machinese Phrase tagger [www.connexor.fi]. The tokens consisted of 266 000 forms of 175 000 base forms.

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, out of which 144 million were morphologically analyzed and disambiguated. The tokens were 4.8 million inflected forms of 1.8 million base forms. 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].

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 morphologically analyzed and disambiguated 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].

For Swahili, we used *The Helsinki Corpus of Swahili* (HCS), which is an annotated corpus of Standard Swahili text. It consisted of 12 million morphologically analyzed and disambiguated running text tokens. The tokens were 268 000 inflected forms of 28 000 base forms. The corpus contains news texts from several current Swahili newspapers as well as from the news site of Deutsche Welle. It also contains extracts from a number of books containing prose text, including fiction, education and sciences. Gatherers are The Department of African and Asian Studies, University of Helsinki. The corpus is available through CSC [www.csc.fi].

Table 1. Corpus data sizes in tokens, types, inflected word forms and base forms (= lexicon look-up forms)

Language	Tokens	Types	Inflected word forms	Base forms	Infl./ Base	Type/ Infl.
Finnish	144 M	3 868 K	4 178 K	1 801 000	2.3	92.3
English	26 M	210 K	222 K	175 000	1.3	94.6
Swedish	35 M	645 K	655 K	445 000	1.5	98.5
Swahili	12 M	231 K	243 K	28 000	8.6	95.1

In Table 1, the difference between the figures in the columns *Types* and *Inflected word forms* means that some word forms have more than one base form, in which case we counted them as separate inflected word forms. This means ambiguity that can only be resolved in context. The ratio between the two gives us an upper-bound on how well any algorithm that only takes the words and not their contexts into account can perform if guessing only the single most likely base form for a new word of the language. By giving several suggestions, the recall can of course go above this figure while the precision goes down.

3.2 Training Data

We have corpora with pairs of word and base forms, for which the correct base form has been mechanically identified in context. We construct our training material by ordering the word and base form pairs according to decreasing frequency and divide the training material into four top frequency ranks as shown in Tables 3a-d.

3.3 Test Data

We draw 5000 word and base form pairs from the frequency rank 100 001-300 000 as test material. The test data frequency ranks can be seen in Table 4.

Table 3a. Top frequency ranks of inflected word and base form pairs for English

Frequency rank	Number of inflected forms	Cum. number of inflected forms	Number of base forms	Cum. number of base forms
1- 3 000	2 720	2 720	2 176	2 176
3 001- 10 000	6 143	8 863	4 566	6 742
10 001- 30 000	16 753	25 616	12 851	19 593
30 001-100 000	54 218	79 834	44 402	63 995

Table 3b. Top frequency ranks of inflected word and base form pairs for Finnish

Frequency rank	Number of inflected forms	Cum. number of inflected forms	Number of base forms	Cum. number of base forms
1- 3 000	2 781	2 781	1 587	1 587
3 001- 10 000	6 346	9 127	2 633	4 220
10 001- 30 000	17 757	26 884	6 601	10 821
30 001-100 000	61 139	88 023	21 301	32 122

Table 3c. Top frequency ranks of inflected word and base form pairs for Swedish

Frequency rank	Number of inflected forms	Cum. number of inflected forms	Number of base forms	Cum. number of base forms
1- 3 000	2 761	2 761	1 898	1 898
3 001- 10 000	6 351	9 112	3 764	5 662
10 001- 30 000	17 593	26 705	9 938	15 600
30 001-100 000	58 629	85 334	33 049	48 649

Table 3d. Top frequency ranks of inflected word and base form pairs for Swahili

Frequency rank	Number of inflected forms	Cum. number of inflected forms	Number of base forms	Cum. number of base forms
1- 3 000	2 619	2 619	2 012	2 012
3 001- 10 000	5 985	8 604	3 208	5 220
10 001- 30 000	17 116	25 720	4 708	9 928
30 001-100 000	60 512	86 232	7 485	17 413

Table 4. Test data frequency rank 100 001-300 000 of inflected word and base form pairs

Frequency rank 100 001-300 000	Number of inflected forms	Number of base forms
English	127 359	111 665
Finnish	170 725	57 306
Swedish	165 929	109 283
Swahili	144 964	10 497

3.4 Baseline and Significance Tests

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. All figures are reported with their 99 % confidence intervals. This means that corresponding test results with non-overlapping confidence intervals are statistically very significantly different.

The baseline assumption is that new words appear in their base form, i.e., we need not do anything. We tested the baseline hypothesis drawing 5000 word and base form pairs at random from the test data frequency rank in Table 4. Since we are only interested in words that we have not seen in the training material, we only count inflected forms of new base forms. As no more than one suggestion is available for each word form in our baseline test, the average baseline precision at maximum recall is identical to the recall in Table 5.

Table 5. Baseline precision and recall for 5000 words drawn from the test data frequency rank

Language	New words (with unseen base form)	New words in base form	Precision & Recall in % ± confidence
English	2912	2508	86.1 ±0.7
Finnish	2081	1051	50.5 ±1.6
Swedish	3395	2043	60.2 ±1.2
Swahili	384	159	41.4 ±3.7

As can be seen from the baseline experiment, around 86 % of the new words in English appear in their base form, whereas the corresponding figures for Swedish is around 60 %, for Finnish around 50 % and for Swahili around 40 %.

4 Experiments

We test how well the analogical guesser is able to predict base forms for new words using the test data for which we calculated the baseline in Section 3.3. The sensitivity of the model is tested using increasing amounts of training data. The model makes no particular assumptions about the language except that the inflections are encoded as prefixes and/or suffixes which may cover parts of the stem if the stem also changes. We test the model on

the new words of the test data using various amounts of training material. The amounts of training data and the corresponding results can be seen in Tables 6a-d.

Table 6a. Recall and average precision in the test frequency rank 100 000-300 000 for English

Training data frequency ranks	Found correct test base forms	Recall in % ± confidence	Avg. precision in % ± confidence
1- 3 000	2734	93.8±0.3	74.9±0.7
1- 10 000	2764	94.8±0.3	78.2±0.6
1- 30 000	2781	95.5±0.2	81.3±0.5
1-100 000	2834	97.5±0.1	88.2±0.4

Table 6b. Recall and average precision in the test frequency rank 100 000-300 000 for Finnish

Training data frequency ranks	Found correct test base forms	Recall in % ± confidence	Avg. precision in % ± confidence
1- 3 000	1964	91.2±0.5	74.2±0.8
1- 10 000	2021	93.9±0.4	78.5±0.7
1- 30 000	2051	95.3±0.3	80.4±0.7
1-100 000	2033	94.4±0.3	79.1±0.7

Table 6c. Recall and average precision in the test frequency rank 100 000-300 000 for Swedish

Training data frequency ranks	Found correct test base forms	Recall in % ± confidence	Avg. precision in % ± confidence
1- 3 000	3294	97.5±0.1	86.2±0.4
1- 10 000	3341	98.9±0.1	88.9±0.3
1- 30 000	3358	99.5±0.05	91.8±0.2
1-100 000	3351	99.2±0.05	94.4±0.2

Table 6d. Recall and average precision in the test frequency rank 100 000-300 000 for Swahili

Training data frequency ranks	Found correct test base forms	Recall in % ± confidence	Avg. precision in % ± confidence
1- 3 000	291	76.0±2.8	69.4±2.8
1- 10 000	320	83.6±2.1	75.7±2.3
1- 30 000	339	89.7±1.4	79.7±1.9
1-100 000	338	89.4±1.5	76.3±2.1

4.1 Importance of the Results

The test results are statistically very highly significant and the test results also indicate that the relative improvements over the baseline are interesting in practice for all four languages as shown in Table 7.

Table 7. Relative improvement over the baseline precision and recall with the maximum amount of training data

Language	Recall	Precision
English	+13.2 %	+2,4 %
Finnish	+86.9 %	+ 56.6 %
Swedish	+64.8 %	+56.8 %
Swahili	+115,9 %	+84.3 %

5 Discussion

In this section, we discuss the test results and give some final notes on the nature of morphologies and the implementation of the model.

We used incremental amounts of training material, i.e. we successively added training material from a new range of ranks while testing on data from the frequency ranks 100 001-300 000. As might be expected, there were successive improvements with additional data. We note, however, that most of the improvements in recall were achieved already with as little training data as the 3 000 most common word and base form pairs of a language and after 10 000 small but significant improvements remained. After the 30 000 most frequent data pairs have been used as training material, the improvements in recall began leveling off. The precision continues to increase for two of the four languages with additional training material. From this, we conclude that using slightly more than the core set of word forms and their corresponding base forms is enough to automatically induce a reasonable guesser for a language. This observation is also true with some caution for human speakers of a language. As an aside, it should be noted that manually editing the most likely base form for each word form in a list of the 30 000 most frequent word forms is a tedious task, but it only takes a week or two for a native linguist.

For most languages, the inflections are affixed to the end or to the beginning of a word stem with some possible minor modification of the stem at the junction. Here Arabic is the most prominently used counter example, for which word inflections are indicated with stem internal vocalization patterns in addition to using affixes. However, the vocalizations are not customarily marked in text except in the Qur'an. In addition, stem changes derive new words, i.e. they are derivational processes of the language not inflectional, e.g., relating words like *book*, *read* and *reader*, and they therefore tend to be lexicalized. However, it remains to be seen whether the model applies as successfully to written Arabic as well.

The model for finding the most likely analog base form for a new word form was implemented with a cascade of weighted finite-state transducers–one for each part of the model. The cascade is composed with the word form at runtime. To extract the most likely base forms, we make a projection of the upper surface of the composed transducer and list the N-best unique base forms, i.e., the N base forms with the smallest total log-probability

weights. The weighted transducers can be implemented in the tropical semiring, where finding the string with the highest probability coincides with the single source shortest distance algorithm. Open Source tools for weighted finite-state transducers have been implemented by, e.g., [6] and [7].

6. Conclusion

We have introduced a new probabilistic model for determining base forms for previously unseen words by analogy with a set word and base form pairs. The model makes no assumptions about whether the inflections are encoded word-initially or word-finally. We tested the model on four morphologically different languages: English, Finnish, Swedish and Swahili. Our model reached a recall of 89-99 % with an average precision of 76-94 % depending on language and the amount of training material. The model was statistically very highly significantly better than the baseline for all four languages and the relative improvement over the baseline was considerable both for recall and precision. From our experiments, it seems like using slightly more than the core set of word forms found in a corpus paired with their base forms would be enough to mechanically induce a reasonable base form guesser, i.e., a lemmatizer for new words of a language.

References

1. Creutz, M., Lagus, K., Lindén, K. and Virpioja, S.:Morfessor and Hutmegs: Unsupervised Morpheme Segmentation for Highly-Inflecting and Compounding Languages. In Proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn, Estonia, April 4-8, (2005).
2. Goldsmith, J.: Morphological Analogy: Only a Beginning. [http://hum.uchicago.edu/~jagoldsm/Papers/analogy.pdf] (2007)
3. Kuenning, G.: Dictionaries for International Ispell. [http://www.lasr.cs.ucla.edu/geoff/ispell-dictionaries.html] (2007)
4. Kurimo, M., Creutz, M., Turunen, V.: Overview of Morpho Challenge in CLEF 2007 In Working Notes of the *CLEF 2007* Workshop. (eds.) Alessandro Nardi and Carol Peters. 19-21 September, Budapest, Hungary. (2007)
5. Lindén, K.: Multilingual Modeling of Cross-lingual Spelling Variants. In Journal of Information Retrieval 9 (2006) 295-310.
6. Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M., OpenFst: A General and Efficient Weighted Finite-State Transducer Library, Lecture Notes in Computer Science, To appear.
7. Lombardy, S., Régis-Gianas,Y., and Sakarovitch, J.,: Introducing Vaucanson**.** Theoretical Computer Science Vol. 328 (2004) 77 – 96.
8. Wicentowski, R.: Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework. PhD Thesis. Baltimore, Maryland (2002)