Making morphologies the "easy" way

Attila Novák[1,2]

[1] MTA-PPKE Hungarian Language Technology Research Group,
[2] Faculty of Information Technology and Bionics
Pázmány Péter Catholic University
50/a Práter street, 1083 Budapest, Hungary
novak.attila@itk.ppke.hu

Abstract. Computational morphologies often consist of a lexicon and some rule component, the creation of which requires various competences and considerable effort. Such a description, on the other hand, makes an easy extension of the morphology with new lexical items possible. Most freely available morphological resources, however, contain no rule component. They are usually based on just a morphological lexicon, containing base forms and some information (often just a paradigm ID) identifying the inflectional paradigm of the word, possibly augmented with some other morphosyntactic features. The aim of the research presented in this paper was to create an algorithm that makes the integration of new words into such resources similarly easy to the way a rule-based morphology can be extended. This is achieved by predicting the correct paradigm for words not present in the lexicon. The supervised machine learning algorithm described in this paper is based on longest matching suffixes and lexical frequency data, and is demonstrated and evaluated for Russian.

Keywords: morphology, paradigm prediction, Russian

1 Introduction

Morphological analysis is an important task in any natural language processing chain, preceding any further analysis of texts. It is also unavoidable in information retrieval, or indexing algorithms, where the lemma of words are to be used in order to have a robust representation of the information present in the documents. In this case, the morphosyntactic features identifying the specific member of the paradigm of the lexical item are irrelevant, only the lemma is required.

Large-scale computational morphologies are usually created using a morphological grammar formalism that minimizes the amount of information necessary to include in the source lexicon about each lexical item by providing some rule-based method of formalization of the morphological behavior of words. This allows an easy extension of the morphology with new lexical items. This approach also gives the creator of the morphology complete control over the quality of the resource. Building rule-based morphological grammars, however, requires threefold competence: familiarity with the formalism, knowledge of the morphology,

phonology and orthography of the language, and extensive lexical knowledge. Many morphological resources, on the other hand, contain no explicit rule component. Such resources are created by converting the information included in some morphological dictionary to some simple data structures representing the inflectional behavior of the lexical items included in the lexicon. The representation often only contains base forms and some information (often just a paradigm ID) identifying the inflectional paradigm of the word, possibly augmented with some other morphosyntactic features. With no rules, the extension of such resources with new lexical items is not such a straightforward task, as it is in the case of rule-based grammars. However, the application of machine learning methods may be able to make up for the lack of a rule component. In this paper, we intend to solve the problem of predicting the appropriate inflectional paradigm of out-of-vocabulary words, which are not included in the morphological lexicon. The method is based on a longest suffix matching model for paradigm identification, and it is showcased with and evaluated against an open-source Russian morphological lexicon.

The context in which we explored the possibilities of automatic paradigm identification, was the following task. We needed to make a pop-up dictionary capable of handling and correctly lemmatizing all inflected word forms of the vocabulary of a specific Russian–Hungarian dictionary. The morphological engine integrated in the dictionary program is Humor ([12, 10]), a constraint-based morphological analyzer, which was first developed for Hungarian morphology. Instead of creating a Humor-based Russian morphology from scratch, we decided to adapt an LGPL-licensed Russian resource, available from www.aot.ru ([13]). The core vocabulary of this morphology is based on Zaliznyak's morphological dictionary [16]. It contains 174 785 lexical entries, each of which are classified into one of 2 767 paradigms. The resource was converted to the Humor formalism, and its coverage needed to be extended to cover the whole vocabulary of the dictionary. For the evaluation of the performance of the paradigm assignment algorithm, we used various disjunct parts of the aot resource. In addition, we used the frequency distribution of Russian lemmas, taken from Serge Sharoff's Russian internet frequency list.[3]

The paper is structured as follows: after a short summary of related work, the features used for predicting Russian inflectional paradigms are described in Section 3. This is followed by description of the suffix model and the ranking algorithm we use. Finally, in Section 6, the performance of the system is evaluated, followed by an error analysis.

2 Related work

Morphological paradigm prediction has been a field of interest, especially for researchers dealing with inflectional, or at least compounding languages. Such languages have a complex morphology, which cannot be covered by hand-made

[3] http://corpus.leeds.ac.uk/frqc/internet-ru.num

lexical resources. Some studies aim at solving this problem by learning inflectional paradigms from raw text corpora by clustering word forms in the corpus and analyzing the resulting clusters ([9, 8, 3]). Other unsupervised methods applied to morphology induction are that of [15], [6] and [5], the latter using morphemes to encode a corpus by grouping morphemes into structures, called signatures, representing inflectional paradigms. These models, however, mainly aim at only segmenting word forms into stems and affixes: stem alternations cause paradigms to be scattered into unrelated subparadigms. However, the performance of unsupervised methods is far behind those using existing resources either as an inventory of inflectional pattern rules, or as annotated data for supervised machine learning algorithms.

Raw text corpora are also used in approaches where word form statistics are used to validate inflectional forms generated by a predicted paradigm candidate for a given word. If the resulting word forms are not represented in a corpus, then the paradigm is not valid. Some examples for such methods are described in [4] and [11]. The algorithm of [7] exploits both lexical features and corpus-based information to determine inflectional behavior by analogy. The author of [14] also defines string-based and corpus-based features used for a support vector machine classifier to decide if a predicted paradigm is valid or not. The most similar approach to our method is the one used in [1], implemented in parallel with our research, however they emphasize paradigm generalization.

Our approach differs from most of the previous ones in that we use a morphological lexicon as annotated data and the frequency distribution of raw text corpora. We address the problem of predicting inflectional paradigms based on the lemma and some given lexical features which are usually available in some less-sophisticated dictionaries. Based on the information coming from the dictionary, the morphological lexicon can be extended in a more robust manner than in cases when only raw word form corpus frequency data is available, and lemma, categorial features and the paradigm all need to be estimated from that data.

3 Features affecting the paradigmatic behavior of Russian words

When attempting to predict the inflectional paradigm for Russian words, certain grammatical features of the lexical item need to be known in order to have a good chance of guessing right. Lemma and part of speech are obviously necessary features, although part of speech can be guessed from the lemma for adjectives and verbs with rather good confidence. Nevertheless, we assumed these to be known, as these properties of words are present in any dictionary.

For nouns, a number of additional features (gender, countability and animacy) play a role in determining the morphosyntatctic feature combination slots which make up the paradigm of the given lemma. There are also nouns, which are undeclinable. Of these features, gender is indicated for each headword in any dictionary, and undeclinable nouns are also usually marked as such. Certain

abstract, collective and mass nouns (and, in the aot resource, also many proper names) lack plural forms, while there are also pluralia tantum, which have no singular. Some of the latter, however, are easier to recognize, due to their lemma exhibiting typical plural morphology.

Animacy affects the nominal paradigm in a manner that does not influence the actual set of possible word forms. However, there is a case syncretism in Russian, which depends on animacy. For animate nouns, plural accusative coincides with genitive (for masculine nouns, the same applies also to singular). For inanimate nouns, on the other hand, the form of accusative matches that of the nominative. This difference is still present in the case of homonyms, where one of the senses of the word is animate, and another form is inanimate. This phenomenon is illustrated in Figure 1 with the word *ёж* 'hedgehog: animal', and 'Czech hedgehog: a static anti-tank obstacle'. Note, however, that the animacy feature, although it is present in the aot lexicon, is not generally made explicit in other dictionaries, because a human user can infer this information from the meaning of the word. We thus have not used this information.

ёж[num:Sg.cas:Nom]	ёж[num:Sg.cas:Nom]
ежа[num:Sg.cas:Gen]	ежа[num:Sg.cas:Gen]
ежу[num:Sg.cas:Dat]	ежу[num:Sg.cas:Dat]
ежа[num:Sg.cas:Acc]	ёж[num:Sg.cas:Acc]
ежом[num:Sg.cas:Ins]	ежом[num:Sg.cas:Ins]
еже[num:Sg.cas:Prp]	еже[num:Sg.cas:Prp]
ежи[num:Pl.cas:Nom]	ежи[num:Pl.cas:Nom]
ежей[num:Pl.cas:Gen]	ежей[num:Pl.cas:Gen]
ежам[num:Pl.cas:Dat]	ежам[num:Pl.cas:Dat]
ежей[num:Pl.cas:Acc]	ежи[num:Pl.cas:Acc]
ежами[num:Pl.cas:Ins]	ежами[num:Pl.cas:Ins]
ежах[num:Pl.cas:Prp]	ежах[num:Pl.cas:Prp]
(a) ёж[N.gnd:Mas.ani:**Ani**][:8];	(b) ёж[N.gnd:Mas.ani:**Ina**][:9];

Fig. 1: Differences in case syncretism of the lemma (*ёж* 'hedgehog') depending on whether it is animate (a) or inanimate (b).

Similarly, the set of valid morphosyntactic feature combinations for verbs depends on verbal aspect and transitivity/reflexivity. Thus, these properties need to be known for verbs, and, indeed, they are listed in dictionaries. E.g. nontransitive verbs lack passive participles; verbs of perfective aspect lack present participle forms; and many verbs of imperfect aspect lack past participial (especially passive) forms. The adverbial participial forms a verb may assume also depend on aspect (and also on other idiosyncratic lexical features).

Defectivities of the adjectival paradigm, e.g. the lack of short predicative forms and synthetic comparative and superlative forms depend on semantic and other, seemingly idiosyncratic, features of the lexeme. E.g. relational adjectives

usually lack these forms. Such properties, however, were not made explicit in the aot lexicon, neither are they present in normal dictionaries, so we did not use any lexical features for adjectives beside part of speech.

Thus, when defining the feature set for predicting inflectional paradigms of words, we assumed that the lemma and the lexical properties mentioned above: part of speech, gender, verb type, etc., are known. Other morphological characteristics relevant for inflection that cannot be derived neither from a simple dictionary, nor from the surface form of a word, such as animacy, optional stress variation, idiosyncratic orthographic variations, or other irregularities were not made available to the system. Thus, our model is not necessarily able to predict paradigmatic behavior depending on such features.

The other set of features we used are n-character-long suffixes of the lemma for various lengths n. The maximum suffix length is a parameter of the algorithm. It was set to 10 in the experiments reported in this paper. In order to exploit this information, a suffix model is created based on the lexicon. An illustration of how this model including both the endings and the lexical features is generated is shown in Figure 2.

4 Creation of the suffix model

A suffix trie is built of words input to the training algorithm in the form shown in the right column of Figure 2.

мумиё[N.n.*.-];prd:25	мумиё|n*[N.n-25]
остриё[N.n.-];sfx:ё;prd:1709	остри#ё|n[N.n-1709]
бабьё[N.n-];sfx:ё;prd:210	бабь#ё|ns[N.n-210]
дубьё[N.n-];sfx:ё;prd:210	дубь#ё|ns[N.n-210]
свежевьё[N.n-];sfx:ё;prd:210	свежевь#ё|ns[N.n-210]
цевьё[N.n.-];sfx:ьё;prd:1433	цев#ьё|n[N.n-1433]
жнивьё[N.n];sfx:ё;prd:1103	жнивь#ё|n[N.n-1103]
суровьё[N.n];sfx:ё;prd:210	суровь#ё|ns[N.n-210]
мостовьё[N.n];sfx:ё;prd:210	мостовь#ё|ns[N.n-210]

Fig. 2: A portion of the suffix model. The format of the right column is: `lem#ma|lex-features[PosTag-paradigmID]`, where `ma` is a required ending of the lemma for all items in the paradigm identified by `paradigmID`.

The lemma is decorated with the following features (from right to left):

- The tag in brackets consists of two parts: part of speech (and, in the example in Figure 2: gender) is followed by the appropriate paradigm ID from the aot database; the two are separated by a hyphen. This is the information to be predicted by the algorithm for unknown words. After processing the training data, terminal nodes of the suffix trie link to a data structure representing the distribution (relative frequency) of tags for the given suffix.

- A suffix following a vertical bar is attached to the end of the lemma. This represents the available lexical knowledge about the lexical item in an encoded form (n: neuter noun, *: undeclinable, s: singular only).
- Some paradigms are restricted to lemmas ending in a specific suffix. There is a hash mark at the beginning of the suffix of the lemma that is required by the given paradigm ID to be valid. The given paradigm ID is not applicable to words not having that ending. E.g. all lemmas in paradigm 1433 must end in *ъё*.

5 Ranking

The suffix-trie-based ranking algorithm that we used was inspired by the suffix guesser algorithm used in Brants' TnT tagger to estimate the lexical probability of out-of-vocabulary words ([2]). However, that model did not prove to perform well enough in this task. So we modified the model step-by-step until we arrived at a model that turned out to be simpler, yet to perform much better. The paradigms are predicted by assigning a score to each paradigm for each word. Then, the higher this score is for a paradigm tag for a certain word, the more probable it is that the word belongs to that paradigm. We select the top-ranked paradigm to be the predicted inflectional class.

The score for each paradigm is calculated for all suffixes of the word, including the lexical properties, from shortest to longest. For all tags, the rank is calculated iteratively according to Formula 1.

$$rank^{i+1}[tag] = sign \times len\_sfx \times rel\_freq + rank^{i}[tag] \tag{1}$$

where

$sign$	is negative if the suffix is shorter than the minimal suffix required by the given paradigm
len_sfx	is the length of suffix not including lexical properties
rel_freq	is the relative frequency of tag for the suffix
$rank^{i}[tag]$	is divided by len_sfx if $len\_sfx > 1$ is negated if $sign > 0$ and $rank^{i}[tag] < 0$ before calculating $rank^{i+1}[tag]$

The applied ranking score clearly prefers the most frequent paradigm for the longest matching suffix. Some examples for the ranked candidates are shown in Figure 3.

6 Evaluation

Evaluation of the ranking algorithm was performed on different training and test set combinations. In each case, we applied five-fold crossvalidation. In order to see how the performance of the algorithms is affected by the frequency of

```
курба|f [N.f]    [N.f:50]#2.857270 [N.f:175]#0.756756 [N.f:48]#0.293840
                 [N.f:105]#0.175658 [N.f:88]#0.098045 [N.f:103]#0.051742
                 [N.f:396]#0.03995 [N.f:611]#0.039730 [N.f:69]#0.029693
                 [N.f:121]#0.021167

дурака|f [N.f]       [N.f:88]#4.466005 [N.f:15]#1.341181 [N.f:273]#0.904291
                     [N.f:36]#0.738748 [N.f:50]#0.467147  [N.f:16]#0.443249
                     [N.f:39]#0.300179 [N.f:105]#0.175658 [N.f:96]#0.155983
                     [N.f:103]#0.051742
```

Fig. 3: The ten highest ranked paradigm candidates for the input words курба|f and дурака|f. The candidates are listed sorted by their rank, with the calculated score separated by the # mark for each tag.

the lemmas in the training and test sets, we split the aot lexicon into parts that contained rare words (LT10; not more than 10 occurrences in the Internet corpus; 91,770 words), average words (LT100; between 11 and 100 occurrences; 33,990 words), and frequent words (MT1000; more than 1000 occurrences; 9,650 words). Moreover, we also evaluated performance on a random 20% sample of the lemmas disregarding frequency (RAND; 159,935 words).

We used standard evaluation metrics for measuring performance. *First-best accuracy* measures the ratio of having the correct paradigm ranked at the first place. This reflects the ability of the system to automatically classify new words to paradigms. In addition, the accuracy values for 2^{nd} to 9^{th} ranks were also calculated. *Recall* is the ratio of having the correct paradigm in the set of the first ten highest ranked candidates. Following the metrics used by [7], precision was calculated as *average precision at maximum recall*, i.e. $1/(1+n)$ for each word, where n is the rank of the correct paradigm. This measures the performance of the ranking algorithm. As it might be the case that paradigm prediction is used to aid human classification, this metric reflects the ratio of noise a human must face with when verifying the results. Finally, *f-measure* is the harmonic mean of precision and recall.

We evaluated our algorithm comparing it to two baseline methods. The first one uses Brants' suffix guesser model ([2]) instead of the longest suffix matching method. This model uses a θ factor to combine tag probability estimates for endings of different length in order to get a smoothed estimate. θ is set as the standard deviation of the probabilities of tags. First, the probability distribution for all suffixes is generated from the training set, then it is smoothed by successive abstraction according to Formula 2.

$$P(t|l_{n-i+1},...l_n) = \frac{\hat{P}(t|l_{n-i+1},...l_n) + \theta_i P(t|l_{n-i},...l_n)}{1+\theta_i} \tag{2}$$

for $i = m...0$, with the initial setting $P(t) = \hat{P}$, where

$\hat{P}$ are maximum likelihood estimates from the frequencies in the lexicon

θ_i weights are the standard deviation of the unconditioned maximum likelihood probabilities of the tags in the training set for all i

The other baseline assigns the most frequent paradigm identifier to each word based on its part of speech and the additional features available (e.g. gender, aspect, etc.). The results of these baselines compared to our system are shown in Table 1. As expected, the second baseline, choosing the most frequent tag, has a rather low accuracy, however, our longest suffix method outperforms the first baseline as well. A key difference between the two models is that Brants' model assigns more weight to unconditioned tag distributions and ones conditioned on shorter suffixes than those conditioned on longer ones. This is just the other way round in the longest suffix algorithm.

Table 1: First-best accuracy of paradigm identifiers achieved by the longest suffix match algorithm, Brants' model, and by assigning the most frequent paradigm tag

	Longest suffix	Brants' model	Most frequent tag
MT1000	0.768	0.587	0.410
LT100	0.876	0.593	0.473
LT10	0.887	0.698	0.480
RAND	0.862	0.632	0.466

The tags containing paradigms ID's as well as detailed PoS and subcategorial features define a very sophisticated classification of words. However, some of the features that distinguish two different paradigms are not relevant from the aspect of their inflectional behavior, such as the subtype of a non-inflecting adverb. Moreover, some of these features cannot even be predicted. In many cases, there is stress variation, which does not affect the set of orthographic forms in the paradigm, however, it yields a different paradigm ID. Moreover, some paradigm differences are irrelevant from the point of view of our dictionary lookup task, because they do not affect the set of word forms in the paradigm. The case syncretism differences between animate and inanimate nouns are examples of such differences. To see how the algorithms perform in our original lemmatization task, equivalence classes of paradigms were generated, and a prediction was considered correct if the set of inflected forms generated by the predicted paradigm was identical to the set of word forms generated by the correct paradigm. Of the 2 767 different paradigms, 921 non-unique paradigms could be collapsed into 283 equivalence classes. Table 2 shows the results for each setup, where rows FULL, ID and EQUIP correspond to full tag, paradigm ID, and equivalence class evaluations, respectively. In the rows marked by ID, instead of full tag agreement, which might include hard-to-predict information like that the word

is the name of an organization, only the paradigm identifiers were considered. Thus `[N.n._nam:Org.--49]`, and `[N.n.--49]` were considered as equivalent.

Table 2: Results on full tag agreement (FULL), paradigm identifiers (ID) and equivalent paradigm classes (EQUIP). The results are measured by first-best accuracy, precision, recall and f-measure.

	MT1000	LT100	LT10	ALL/MT1000	ALL/LT100	ALL/LT10	RAND
FULL	0.752	0.849	**0.879**	0.759	0.855	0.872	0.848
	0.819	0.903	0.926	0.823	0.910	0.923	0.903
	0.903	0.979	0.991	0.923	0.989	0.994	0.982
	0.859	0.940	**0.958**	0.870	0.948	0.957	0.941
ID	0.768	0.876	**0.887**	0.771	0.872	0.885	0.862
	0.830	0.920	0.934	0.834	0.924	0.933	0.915
	0.905	0.980	0.992	0.926	0.990	0.994	0.983
	0.866	0.949	**0.962**	0.878	0.956	0.962	0.948
EQUIP	0.819	0.889	**0.892**	0.813	0.884	0.890	0.875
	0.869	0.929	0.937	0.866	0.932	0.936	0.924
	0.929	0.984	0.993	0.951	0.993	0.995	0.988
	0.898	0.956	**0.964**	0.907	0.961	0.965	0.955

The three columns on the left show results where the models were trained only on words in the same frequency class they were tested on. The test set was always 20% of the lemmas in the given frequency range. Results in the next four columns were obtained by training the models on the complement of the test set w.r.t. the whole lexicon.

As the numbers show, our system performs best on rare words, while it achieved the worst results on very frequent words. This is not very surprising, as irregular words tend to be frequent words, while rare words have regular inflectional behavior. Correctly predicting the exact paradigm of an unknown personal pronoun or an irregular verb is indeed a rather difficult task. Since our aim was to extend existing morphological lexicons, and such resources already contain the most frequent words of the language, the results obtained for rare words are the ones which are relevant for our task.

Also note that beside similar recall values, precision and first-best accuracy are higher when equivalent paradigms are collapsed. The prediction algorithm works reasonably well for extending resources for tasks that do not require full morphological analysis such as indexing for information retrieval or dictionary lookup.

Table 3 shows the first-best paradigm ID accuracy results for all words, nouns, verbs and adjectives separately. The exact paradigm of verbs and adjectives turned out to be more difficult to guess than that of nouns. The results achieved

for adjectives seem to be especially contradictory to the overall performance, which can be explained by the unpredictable behavior of adjectives. Semantic factors and hard-to-predict stress variation affecting paradigmatic classification are explained in the next section of this paper.

Table 3: First-best accuracy of paradigm ID prediction in the case of all types of words, nouns, verbs and adjectives

	ALL	NOUNS	VERBS	ADJECTIVES
MT1000	0.768	0.814	0.702	0.683
LT100	0.876	0.935	0.802	0.772
LT10	0.887	0.968	0.869	0.732
RAND	0.862	0.947	0.848	0.682

7 Error analysis

The most frequent confusions of the longest suffix algorithm for infrequent words are due to failure to correctly predict

- whether an adjective has synthetic comparative, superlative and/or short predicative forms
- whether a *-ние*-final abstract noun has an alternative *-нье* spelling
- whether a noun has a second genitive (used in partitive constructions) or locative form
- stress in past passive participles of certain verb classes and in short and comparative forms of certain adjectives, or other optional stress variation across the paradigm (this results in an *e*∼*ё* contrast not normally reflected in orthography)
- whether a non-inflecting noun can be interpreted as plural
- whether an imperfective verb has past passive participle forms

Except for stress-related issues and semantically motivated or idiosyncratic defectivity, incorrect forms are very rarely predicted by the algorithm. Humans would probably make similar mistakes for words they do not know, especially if they do not know the meaning of the word either. The system sometimes highlights inconsistencies in the original aot data that even the author of this article, who is not a native or even advanced speaker of Russian, can identify as errors, e.g. that while the name of the energy company *Кубаньэнерго* is categorized as lexically non-plural, the similarly formed *Сахалинэнерго* does not have this property.

When looking at errors the algorithm makes when applied to frequent words, we find that the types of errors are similar. Nevertheless, failure to predict superlatives, comparatives, second genitives or special locative forms is more prevalent for this data, as a much higher proportion of very frequent words have these "irregular" forms.

The most frequent errors of Brants' original suffix guesser algorithm, on the other hand, include absurd errors that would not be made even by beginning learners of Russian. This is due to overemphasizing distributions conditioned on shorter suffixes over those on longer ones. The top-ranked candidate paradigm is often totally inapplicable to words having the ending the given lexical item has, such as the paradigm of *-кий*-final adjectives to *-ный*-final ones (the most frequent error of that algorithm for infrequent words).

8 Conclusion

In this article, we presented and evaluated a suffix-trie-based supervised learning algorithm capable of predicting inflectional paradigms for words based on the ending of their lemma and some basic lexical properties. The algorithm can be used to automatically extend the vocabulary of computational morphologies lacking an independent rule component, which is often the case for resources based on a morphological dictionary. The experiments were demonstrated for Russian, however, with minimal adaptation the tool can be used for any language provided there is a morphological resource available. Moreover, we assumed that a dictionary with some lexical features is also available, thus such features could be used for disambiguating paradigm candidates. The results showed that our method can correctly identify the paradigm of unseen words with an accuracy of about 90%, achieving the best performance on relatively rare words, which are good candidates of being absent in the original lexicon. For rare nouns, the paradigm identification accuracy is 96.8%.

We found that assigning more weight to distributions conditioned on longer suffixes than on shorter ones yields much better prediction performance, not only in terms of the number of exact predicted paradigm matches, but especially when taking into account what sorts of errors the system makes. While the baseline suffix guesser algorithm often proposes paradigms inapplicable to the given lexical item, our algorithm makes errors that arise due to the lack of lexical semantic information. Humans would make similar errors in similar situations.

Acknowledgement

The author of this article would like to thank Borbála Siklósi for her help, especially in evaluation.

References

1. Ahlberg, M., Forsberg, M., Hulden, M.: Semi-supervised learning of morphological paradigms and lexicons. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden 26–30 April 2014. pp. 569–578 (2014), http://aclweb.org/anthology//E/E14/E14-1060.pdf

2. Brants, T.: Tnt - a statistical part-of-speech tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000). Seattle, WA (2000)
3. Dreyer, M., Eisner, J.: Discovering morphological paradigms from plain text using a dirichlet process mixture model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 616–627. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
4. Forsberg, M., Hammarström, H., Ranta, A.: Morphological lexicon extraction from raw text data. In: Proceedings of the 5th International Conference on Advances in Natural Language Processing. pp. 488–499. FinTAL'06, Springer-Verlag, Berlin, Heidelberg (2006)
5. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Comput. Linguist. 27(2), 153–198 (Jun 2001)
6. Hammarström, H., Borin, L.: Unsupervised learning of morphology. Comput. Linguist. 37(2), 309–350 (Jun 2011)
7. Linden, K.: Entry generation by analogy – encoding new words for morphological lexicons. In: Journal Northern European Journal of Language Technology. pp. 1–25 (2009)
8. Monson, C., Carbonell, J.G., Lavie, A., Levin, L.S.: Paramor: Finding paradigms across morphology. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF. Lecture Notes in Computer Science, vol. 5152, pp. 900–907. Springer (2007)
9. Nakov, P., Bonev, Y., Angelova, G., Gius, E., von Hahn, W.: Guessing morphological classes of unknown german nouns. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) RANLP. Current Issues in Linguistic Theory (CILT), vol. 260, pp. 347–356. John Benjamins, Amsterdam/Philadelphia (2003)
10. Novák, A.: What is good Humor like? [Milyen a jó Humor?]. In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)
11. Oliver, A., Tadic, M.: Enlarging the croatian morphological lexicon by automatic lexical acquisition from raw corpora. In: LREC. European Language Resources Association (2004)
12. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 261–268. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)
13. Sokirko, A.V.: Morphological modules at the site www.aot.ru. In: Dialog'2004 (2004)
14. Šnajder, J.: Models for predicting the inflectional paradigm of croatian words. In: Slovenščina 2.0. pp. 1–34 (2013)
15. Wicentowski, R.: Modeling and learning multilingual inflectional morphology in a minimally supervised framework. Tech. rep. (2002)
16. Zaliznyak, A.A.: Russian grammatical dictionary – Inflection. Russkij Jazyk, Moskva (1980)