# Portable Language Technology: Russian via Czech

Jiri Hana
Department of Linguistics
The Ohio State University
Columbus, OH 43210
hana@ling.osu.edu

Anna Feldman
Department of Linguistics
The Ohio State University
Columbus, OH 43210
afeldman@ling.osu.edu

**Abstract**

We report on morphological tagging of Russian using very limited Russian resources. We train the TnT tagger (Brants, 2000) on a modified Czech corpus to get the transition probabilities. We believe that the two languages are similar enough for the transitional information to be useful. The Russian emission symbols are obtained using a morphological analyzer that does not rely on a manually created lexicon. Finally, we report on several simple systematic modifications transforming a Czech text into a text with more Russian-like morphological properties.

## 1 Introduction

Morphological processing and tagging is essential for many NLP tasks, including machine translation, information retrieval and parsing. In this paper, we describe an automatic morphological analysis and tagging of Russian eschewing the use of extensive resources; particularly, large annotated corpora and lexicons. Performing such an analysis is not trivial, because Russian is a highly inflective language with a high degree of morpheme homonymy (cf. Table 1)[1]:

| | |
|---|---|
| krasiv-a | beautiful (short adjective, feminine) |
| muž-a | table (noun, masc., sing., genitive) |
| | table (noun, masc., sing., accusative) |
| okn-a | window (noun, neuter, sing., genitive) |
| | window (noun, neuter, plural, nominative) |
| | window (noun, neuter, plural, accusative) |
| knig-a | book (noun, fem., sing., nominative) |
| dom-a | house (noun, masc., sing., genitive) |
| | house (noun, masc., plural, nominative) |
| | house (noun, masc., plural, accusative) |
| skazal-a | say (verb, fem., sing., past tense) |
| dv-a | two (numeral, masc., nominative) |

**Table 1: Homonymy of the *a* ending**

Since there is no morphologically annotated large-scale corpus freely available for Russian, we could not employ the standard methods used in stochastic POS tagging. Instead, we decided to exploit the existing large annotated corpora of Czech, a genetically related language with similar linguistic properties (free word order and rich morphology which play a great role in determining agreement and argument relationships).

We trained the TnT tagger (Brants, 2000) on Czech to get the transition probabilities and we performed the morphological analysis of Russian to get the emission symbols for TnT. Given that both Russian and Czech have relatively free word order, it may seem an odd choice to use a Markov model tagger. Why should second order Markov models be able to capture useful facts about such languages? Firstly, at a theoretical level, even if a language has the potential for free word order, it may still be that there are recurring patterns in the progressions of parts-of-speech that are attested in a training corpus. Secondly, at a practical level, this seems to be the case: we tested TnT on the task of using the Czech corpus to tag Czech, and found performance close to the best available. We can therefore assume that the information captured by the second-order Markov model is useful for Czech (the language from which it was acquired). The present paper shows that transitional information acquired from Czech is also useful for Russian.

We tagged Russian with the created model and evaluated the results. For evaluation, we manually morphologically annotated a small portion of a Russian corpus: the translations of Orwell's "1984" of the MULTEXT-EAST project (Véronis, 1996).

[1]All Russian examples in this paper are transcribed in Roman alphabet. Our system is able to analyze Russian text in both Cyrillic and various transcriptions.

## 2 Russian morphology

Russian morphology is quite complex. Nouns and adjectives distinguish 3 genders, 2 numbers and 6 cases; verbs inflect for number, person and tense; participles for gender and number; adjectives and many adverbs distinguish grade; etc. There are 4 declension classes of nouns and 3 conjugation classes of verbs; each declension class has several paradigms, and for each paradigm, there are some subparadigms, and so on.

Moreover there are phonological and orthographic alternations. By phonological alternations, we mean cases, such as *pi**s**at'* 'write.INF' vs. *pi**š**u* 'write.1SG.Present' or *postri**č**'sja* 'have-haircut.INF.perfective' vs. *postri**ž**etsja* 'have-haircut.3SG.Future'. The **s** and **š** or **č** and **ž** in these examples belong to the morphological stems *pis-* and *-strič-*, respectively.

## 3 Russian versus Czech

A deep comparative analysis of Czech and Russian is far beyond the scope of this paper. However, we would like to mention just a number of the most important facts. Both languages are Slavic (Czech is West Slavonic, Russian is East Slavonic). Both have extensive morphology whose role is important in determining the grammatical functions of phrases. In both languages, the main verb agrees in person and number with subject; adjectives agree in gender, number and case with nouns. Both languages are free constituent order languages. The word order in a sentence is determined by mainly by discourse. It turns out that the word order in Czech and Russian is very similar. For instance, old information mostly precedes new information. The "neutral" order in the two languages is Subject-Verb-Object. Here is a parallel Czech-Russian example from our development corpus:

(1) a. [Czech]
Byl jasný,
was$_{Masc.Past}$ bright$_{Masc.Sg.Nom}$
studený dubnový
cold$_{Masc.Sg.Nom}$ April$_{Masc.Sg.Nom}$
den i hodiny
day$_{Masc.Sg.Nom}$ and clocks$_{Fem.Pl.Nom}$
odbíjely třináctou.
stroke$_{Fem.Pl.Past}$ thirteenth$_{Fem.Sg.Acc}$

b. [Russian]
Byl jasnyj,
was$_{Masc.Past}$ bright$_{Masc.Sg.Nom}$
xolodnyj aprel'skij
cold$_{Masc.Sg.Nom}$ April$_{Masc.Sg.Nom}$
den' i časy
day$_{Masc.Sg.Nom}$ and clocks$_{Pl.Nom}$
probili trinadtsat'.
stroke$_{Pl.Past}$ thirteen$_{Acc}$

'It was a bright cold day in April, and the clocks were striking thirteen.' [from Orwell's '1984']

Of course, not all utterances are so similar. Section 5.4 mentions examples of some systematic differences.

## 4 Realization

### 4.1 The tag system

We adopted the Czech tag system (Hajič, 2000) for Russian, since Russian is very similar to Czech in many linguistic properties. Every tag is represented as a string of 15 symbols each corresponding to one morphological category. For example, the word *vidjela* is assigned the tag VpFS- - -XR-AA - - -, because it is a verb (V), past participle (p), feminine (F), singular (S), does not distinguish case (-), possessor's gender (-), possessor's number (-), can be any person (X), is past tense (R), is not gradable (-), affirmative (A), active voice (A), and does not have any stylistic variants (the last hyphen).

| No. | Description | No. of values | |
|---|---|---|---|
| | | Cz | Ru |
| 1 | POS | 12 | 12 |
| 2 | SubPOS – detailed POS | 75 | 32 |
| 3 | Gender | 11 | 5 |
| 4 | Number | 6 | 4 |
| 5 | Case | 9 | 8 |
| 6 | Possessor's Gender | 5 | 4 |
| 7 | Possessor's Number | 3 | 3 |
| 8 | Person | 5 | 5 |
| 9 | Tense | 5 | 5 |
| 10 | Degree of comparison | 4 | 4 |
| 11 | Negation | 3 | 3 |
| 12 | Voice | 3 | 3 |
| 13 | Unused | 1 | 1 |
| 14 | Unused | 1 | 1 |
| 15 | Variant, Style | 10 | 2 |

**Table 2: Overview and comparison of the tagsets**

The tagset used for Czech (4290 tags) is larger than the tagset we use for Russian (900 tags). There is a theoretical reason for that – Russian morphological categories have in general fewer values (e.g., 6 cases in Russian vs. 7 in Czech; Czech has often formal and colloquial variants of the same morpheme). However, there is also a practical reason – the Czech tag system is very elaborate, while our tagset captures only the core of Russian morphology. Still, the tagset is much larger than the Penn Treebank tagset, which uses only 45 tags (Marcus et al., 1993). Note that a large tagset does not necessarily imply a more complicated task (Elworthy, 1995).

### 4.2 Morphological analysis

In order to get the emissions for tagging Russian with the TnT tagger, we implemented a morphological anal-

ysis of Russian. Our morphological analyzer uses very little manually created data.

#### 4.2.1 Paradigms

The most important feature of our mophological analyzer is that, unlike the morphological analyzers that exist for Russian (Segalovich and Titov, 2000; Segalovich, 2003; Segalovich and Maslov, 1989; Kovalev, 2002; Mikheev and Liubushkina, 1995; Yablonsky, 1999, among others), it does not rely on a manually created lexicon. The system uses a list of morphological paradigms, a short list inflectional prefixes (negative *ne* and superlative *nai*) and a list of closed class terms. This is in keeping with our aim of being resource-light. The system should be therefore relatively easy to port to any other Slavic language.

Our database contains 80 paradigms. We just encode textbook facts about the Russian morphology (cf. Wade, 1992), excluding the majority of exceptions. A paradigm is a set of endings together with the tags that can go with a particular set of stems. Thus, for example, the paradigm below is a set of inflections that go with the masculine stems ending on the "hard" consonants. The tag system we used for this project is described in section 4.1.

| | |
|---|---|
| 0 | NNMS1 - - - - - - - - - - |
| a | NNMS2 - - - - - - - - - - |
| u | NNMS3 - - - - - - - - - - |
| a | NNMS4 - - - - - - - - - - |
| u | NNMS4 - - - - - - - - - 1 |
| e | NNMS6 - - - - - - - - - - |
| u | NNMS6 - - - - - - - - - 1 |
| om | NNMS7 - - - - - - - - - - |
| y | NNMP1 - - - - - - - - - - |
| ov | NNMP2 - - - - - - - - - - |
| am | NNMP3 - - - - - - - - - - |
| ov | NNMP4 - - - - - - - - - - |
| ax | NNMP6 - - - - - - - - - - |
| ami | NNMP7 - - - - - - - - - - |

**A paradigm for masculine nouns that end on the "hard" consonants; e.g., *slon* ('elephant'), *stol* ('table')**

#### 4.2.2 Phonology

We use the terms *ending* and *stem* in a rather technical way, and they do not correspond exactly to the traditional linguistic terms. A stem is the part of the word that does not change within its paradigm; the ending is the part of the word following such a stem. For example, the forms of the verb *moč'* 'can.INF': *mogu* '1sg', *možeš'* '2sg', *možet* '3sg', etc. are analyzed as the stem *mo* followed by the endings *gu*, *žeš'*, *žet*. A more linguistically oriented analysis would involve the endings *u*, *eš'*, *et* and phonological alternations in the stem. All stem internal variations are treated as suppletions – such words belong to different stems of a single lemma. However, such a specification of the paradigms is used by the morphological analyzer only internally. We use a compiler that produces it from a specification, linguistically more plausible. It is possible to specify basic paradigms and then describe exceptions and phonological alternations. Moreover, similar paradigms can be related by inheritance. This approach is similar to (Mikheev and Liubushkina, 1995) and (Hajič, 2004).

#### 4.2.3 Procedure

When analyzing a word, the basic morphological analysis first checks a list of monomorphemic closed-class words and then segments the word into all possible prefix-stem-ending triples. The result has quite high recall (95.4%), but the average ambiguity is very high (10.9 tags/token), and even higher for open class words. We use two approaches to reduce the ambiguity – a longest ending filtering and an automatically acquired lexicon of stems.

#### 4.2.4 Longest ending filtering (LEF)

The first approach is based on a simple heuristic – the correct ending is usually one of the longest candidate endings. In English, it would mean that if a word is analyzed either as having a zero ending or an *-ing* ending, we would consider only the latter; obviously, in the vast majority of cases that would be the correct analysis. In addition, we specify that a few long but very rare endings should not be included in the maximum length calculation (e.g., 2 person pl. imperative).

#### 4.2.5 Deriving a lexicon

The second approach uses a large raw corpus[2] to acquire an open class lexicon of possible stems with their paradigms. It is based on the idea that open-class lemmata are likely to occur in more than one form. First, the morphological analyzer is run on the text (without any filtering), then the entries that occurred with at least certain number of distinct forms and cover the highest number of forms are added to the lexicon. For example, if $\{f_1, f_2\}$ can be analysed as forms of the lemma $l_1$; $\{f_1, f_2, f_3\}$ as $l_2$, and $\{f_2, f_3, f_4\}$ as $l_3$, then we would add $l_2$ and $l_3$ to the lexicon. If the word *talking* is encountered, using the information about paradigms, we can assume that it is either the *-ing* form of the lemma *talk* or that it is a monomorphemic word (such as *sibling*). Based on this single form we cannot really say more. However, if we also encounter the forms *talk*, *talks* and *talked*, the former analysis seems more probable; and therefore, it is reasonable to include the lemma *talk* as a verb into the lexicon. If we encountered also

[2]We used The Uppsala Russian Corpus (1M tokens), which is freely available from Uppsala University at `http://www.slaviska.uu.se/ryska/corpus.html`.

*talkings*, *talkinged* and *talkinging*, we would include both lemmata *talk* and *talking* as verbs. All forms can be treated as equal or they can be weighted. For example, if there are two competing lemmata with the same number of forms, but one occurs only in rare forms (e.g., transgressives), while the other occurs in frequent forms (e.g., nominative), the latter is more likely to be correct. Hana (2004) reports experiments with various weights derived from the Czech corpus, but it turns out that the uniform distributions produce better results (at least on our development corpus). Also, the best results are obtained when the minimal number of distinct forms for a lemma to be considered was set to two.

Obviously, the morphological analysis based on such a lexicon overgenerates, but it overgenerates much less than the analysis based on the endings only. For example, for the word form *partii* of the lemma *partija* 'party', the analysis with a lexicon generated from 1M-word corpus gives 8 possibilities – the 5 correct ones (noun fem sg gen/dat/loc sg and pl nom/acc) and 3 incorrect ones (noun masc sg loc, pl nom, and noun neut pl acc; note that only gender is incorrect). The analysis based on endings only gives 20 possibilities – 15 incorrect (including adjectives and an imperative).

### 4.3 Tagging

We use the TnT tagger (Brants, 2000), an implementation of the Viterbi algorithm for second order Markov models. We train the transition probabilities on Czech (1.5M tokens of the Prague Dependency Treebank (Bémová et al., 1999)). For each set of testing data, we obtain the emission probabilities as uniform distribution of tags given by our morphological analyzer.

## 5 Experiments

### 5.1 Corpora

For evaluation purposes, we selected and manually morphologically annotated a small portion from the Russian translation of Orwell's '1984'. This corpus contains 4011 tokens and 1858 types. During the development, we used another part of '1984'. Since we want to work with minimal language resources, the development corpus is intentionally small – 1788 tokens. We used it to test our hypotheses and tune the parameters of our tools.

In the following sections, we discuss our experiments and report the results. Note that we do not report the results for tag position 13 and 14, since these positions are unused; and therefore, are always trivially correct.

### 5.2 Morphological analysis

As can be seen from Table 4, morphological analysis without any filters gives a good recall; although on a non-fiction text it would be probably lower), but also very high average ambiguity. Both filters (the longest-ending filtering and automatically acquired lexicon) lower the ambiguity significantly; the former with a considerable drop of recall, while the latter retains high recall. However, their combination keeps the recall reasonably high while decreasing the ambiguity the most. As expected, the lexicon acquired on the larger corpus gives better results.

### 5.3 Tagging

In Table 4, we report the results of our experiments. Our baseline is produced by the morphological analyzer without any filters followed by a tagger randomly selecting a tag among the tags offered by the morphological analyzer. The rest of the experiments use different combinations of the longest ending filtering and different sizes of the lexicon. Tagging in combination with the longest ending and the lexicon filters gives us significantly better results than tagging without them. The numbers we obtain are worse than the numbers reported for Czech (Hajič et al., 2001) (95.16% accuracy). However, they use an extensive manually created morphological lexicon (200K+ entries) which gives 100.0% recall on their testing data. Moreover, they train and test their taggers on the same language. To our knowledge, our present results are not directly comparable with anything reported for Russian.

### 5.4 "Russification"

We also experimented with "russified" models. We trained the TnT tagger on the Czech corpus with modifications that made the training data look more like Russian. For example, Czech and Russian verbs correspond quite well in their classes and conjugational patterns, though there are some differences. Conditional and "subjunctive" meaning is expressed in Czech by forms in *by-* conjugated for all three numbers and both persons, together with the past participle, whereas in Russian only the particle *by* is used for all numbers and persons, also with past participle.

(2)

| Czech | Russian | |
|---|---|---|
| Já **bych** spal. | Ja **by** spal. | 'I would sleep.' |
| Ty **bys** spal. | Ty **by** spal. | 'You.sg would sleep.' |
| On **by** spal. | On **by** spal. | 'He would sleep.' |

Plural adjectives and participles in Russian, unlike Czech, do not distinguish gender.

(3) a. Nadaní sportovci zpívali
Gifted$_{Masc.Pl}$ sportsmen sang$_{Masc.Pl}$
vlastenecké písně.
patriotic songs
'Gifted sportsmen were singing patriotic songs.' [Cz]

b. Nadané sportovkyně zpívaly
Gifted$_{Fem.Pl}$ sportswomen sang$_{Fem.Pl}$

vlastenecké písně.
patriotic songs
'Gifted sportswomen were singing patriotic songs.' [Cz]

c. Nadaná děvčata zpívala
Gifted$_{Neut.Pl}$ girls$_{Neut}$ sang$_{Neut.Pl}$
vlastenecké písně.
patriotic songs
'Gifted girls were singing patriotic songs.' [Cz]

d. Talantlivye sportsmeny/sporstmenki peli
Gifted$_{Pl}$ sportsmen/sportswomen sang$_{Pl}$
patriotičeskie pesni.
patriotic songs
'Gifted sportsmen/sportswomen were singing patriotic songs.' [Ru]

Negation in Czech is in the majority of cases is expressed by the prefix *ne-*, whereas in Russian it is very common to see a separate particle (*ne*) instead:

(4) a. Nic **nedělal**.
nothing not-did
'He didn't do anything.' [Cz]

b. On ničego **ne delal**.
he nothing not did
'He didn't do anything.' [Ru]

In addition, reflexive verbs in Czech are formed by a verb followed by a reflexive clitic, whereas in Russian, the reflexivization is the affixation process:

(5) a. Filip **se** ještě neholí.
Filip REFL-CL still not-shaves
'Filip doesn't shave yet.' [Cz]

b. Filip esče ne breet+**sja**.
Filip still not shaves+REFL.SUFFIX
'Filip doesn't shave yet.' [Ru]

The present tense copula is obligatory in Czech, whereas in Russian, its use is only for emphasis:

(6) a. Já **jsem** psal.
I aux$_{1sg}$ wrote
'I was writing/I wrote.' [Cz]

b. Ja pisal.
I wrote
'I was writing/I wrote.' [Ru]

We implemented a number of "russifications" and some of them are summarized in Table 3. The bold scores indicate that the performance of a russified tagger is better than the original. We admit that we expected more significant gains in accuracy.

The random omission of the copula worsens the overall performance of the tagger. However, its combination with the omission of the reflexive clitics improves the overall results by 1.3%; and its combination with the "russified" negation improves the results by 1%. The "russified" negation model and the "russified" reflexive clitics model on their own improve the overall performance, as well as their combinations, but the most significant improvement is obtained when all the three "russifications" are combined together.

## 6 Ongoing Research

We are currently working on improving both the morphological analysis and tagging. We would like to improve recall of filters following morphological analysis, e.g., using n maximal values instead of 1, using some basic knowledge of derivational morphology, etc. We are incorporating phonological conditions on stems into the guesser module as well as are trying to deal with different morphological phenomena specific to Russian, e.g., verb reflexivization. However, we try to stay as much as possible language independent (at least within Slavic languages) and limit the language dependent parameters to an absolute minimum.

We are currently running a set of experiments that involves training TnT on sub-parts of the tag as well as on different combinations of the tag slots and combining the resulting models by simple majority voting and best first unpacking. The combined model outperforms the best model cited in this paper by 2.2%. The details of the experiments are reported in (Hana et al., 2004). In addition to the classifiers that are based on the same learning strategy but trained on different training set, we are working on combining different types of taggers, especially models with a different bias, such as a transformation-based learner, or a discriminative learner such as a maximum entropy tagger.

If possible, we would like to avoid entirely throwing away the Czech emission probabilities, because our intuition tells us that there are useful lexical similarities between Russian and Czech, and that some suitable process of cognate detection will allow us to transfer some information from the Czech to the Russian emission probabilities. We are seeking for a sufficiently general algorithm to make the method portable to other languages, for which we assume we have neither the time nor the expertise to undertake knowledge-intensive work. A suitably automatic algorithm is described by (Kondrak, 2001).

Finally, we would like to extend our work to Slavic languages for which there are even fewer available resources than Russian, such as Belarusian or Ukranian, with the aim of better understanding the portability implications of our approach.

## 7 Acknowledgements

We would like to thank to Chris Brew for invaluable comments and to Jan Hajič for providing assistance with the morphological analysis and Czech tag-system.

| | non-russified | c | n | r | cr | cn | rn | crn |
|---|---|---|---|---|---|---|---|---|
| overall | 68.0 | 67.9 | **69.0** | **69.4** | **69.3** | **69.0** | **69.4** | ***69.5*** |
| 1 POS | 87.9 | 87.8 | **89.0** | **88.8** | **88.8** | **89.0** | **89.0** | **89.0** |
| 2 SubPOS | 86.0 | 86.0 | **86.5** | **86.5** | **86.4** | **86.5** | **86.6** | **86.6** |
| 3 Gender | 80.9 | 80.7 | **81.3** | **81.3** | **81.3** | **81.2** | **81.4** | **81.4** |
| 4 Number | 91.5 | **91.6** | **92.2** | **92.1** | **92.2** | **92.3** | **92.3** | **92.4** |
| 5 Case | 80.5 | **80.8** | **80.6** | **80.9** | **81.0** | **80.8** | **80.9** | **80.9** |
| 6 PossGender | 98.5 | 98.5 | 98.4 | 98.4 | 98.3 | 98.5 | 98.4 | 98.5 |
| 7 PossNumber | 99.6 | 99.7 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 |
| 8 Person | 98.7 | 98.7 | 98.2 | 98.3 | 98.3 | 98.2 | 98.3 | 98.3 |
| 9 Tense | 96.8 | 96.8 | **97.1** | **97.1** | **97.1** | **97.1** | **97.0** | **97.0** |
| 10 Grade | 96.8 | 95.2 | 96.0 | 95.9 | 95.9 | 96.0 | 96.0 | 96.0 |
| 11 Negation | 96.8 | 96.6 | **97.0** | 96.8 | 96.6 | **97.0** | 96.9 | **97.0** |

**Table 3: Performance of the TnT tagger on the "russified" data (development corpus)**
**c = random omission of the copula; n = "russified" negation; r = omission of the reflexive clitics**

| Longest Ending Filtering | no | no | no | yes | yes | yes | yes | Baseline | Unigram |
|---|---|---|---|---|---|---|---|---|---|
| L2 trained on | 0 | 100K | 1M | 0 | 100K | 1M | 1M | | Entropy |
| "russified" | | | | | | | yes | | |
| MA | | | | | | | | | - |
| recall | **95.4** | 94 | 93.1 | 84.4 | 88.3 | 90.4 | 90.4 | **95.4** | - |
| avg. ambiguity (tags/word) | 10.9 | 7.0 | 4.7 | 4.1 | 3.5 | **3.1** | **3.1** | 10.9 | - |
| Tagging – accuracy | | | | | | | | | |
| Tags | 50.7 | 62.1 | 67.5 | 62.1 | 66.8 | 69.4 | **72.6** | 33.6 | |
| Position 1 (POS) | 74.2 | 82.3 | 86.5 | 83.5 | 86.7 | 88.5 | 90.1 | 63.2 | 3.02 |
| Position 2 (SubPOS) | 71.4 | 80.3 | 85.1 | 80.6 | 84.3 | 86.8 | 88.2 | 57.0 | 3.60 |
| Position 3 (Gender) | 70.7 | 77.6 | 81.6 | 78.1 | 80.6 | 82.5 | 84.5 | 59.2 | 2.06 |
| Position 4 (Number) | 84.3 | 88.1 | 90.3 | 89.4 | 90.7 | 91.2 | 92.6 | 75.9 | 1.48 |
| Position 5 (Case) | 60.8 | 72.7 | 78.0 | 76.9 | 79.5 | 80.4 | 84.2 | 47.3 | 2.29 |
| Position 6 (PossGen) | 90.6 | 94.6 | 97.1 | 98.5 | 98.0 | 98.4 | 98.8 | 83.4 | 0.45 |
| Position 7 (PossNr) | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 0.43 |
| Position 8 (Person) | 98.6 | 99.2 | 99.6 | 97.8 | 98.8 | 99.3 | 98.9 | 97.1 | 0.35 |
| Position 9 (Tense) | 90.9 | 94.3 | 95.7 | 95.9 | 96.7 | 96.5 | 97.6 | 86.6 | 0.55 |
| Position 10 (Grade) | 92.3 | 94.6 | 96.1 | 92.0 | 94.0 | 95.9 | 96.6 | 90.1 | 0.47 |
| Position 11 (Neg) | 88.0 | 91.9 | 94.4 | 93.9 | 94.7 | 95.3 | 95.5 | 81.4 | 1.04 |
| Position 12 (Voice) | 90.9 | 94.6 | 96.2 | 96.7 | 97.4 | 97.2 | 97.9 | 86.4 | 0.49 |

**Table 4: Comparison of morphological analysis and tagging with various parameters (test corpus)**

**References**


Bémová, A., J. Hajič, B. Hladká, and J. Panevová (1999). Morphological and Syntactic Tagging of the Prague Dependency Treebank. In *Proceedings of ATALA Workshop*, pp. 21–29. Paris, France.

Brants, T. (2000). TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP-NAACL*, pp. 224–231.

Elworthy, D. (1995, April). Tagset design and inflected languages. In *EACL SIGDAT workshop "From Texts to Tags: Issues in Multilingual Language Analysis"*, Dublin, pp. 1–10.

Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. In *Proceedings of ANLP-NAACL Conference*, Seattle, Washington, USA, pp. 94–101.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. In press.

Hajič, J., P. Krbec, P. Květoň, K. Oliva, and V. Petkevič (2001). Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In *Proceedings of ACL Conference*, Toulouse, France.

Hana, J. (2004). Automatic Lexical Acquisition from Russian Raw Corpora. Forthcoming.

Hana, J., A. Feldman, and C. Brew (2004). A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. Submitted to EMNLP 2004.

Kondrak, G. (2001, June). Identifying cognates by phonetic and semantic similarity. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, pp. 103–110.

Kovalev, A. (2002). A Probabilistic Morphological Analyzer for Russian and Ukranian. http://linguist.nm.ru/stemka/stemka.html (in Russian).

Marcus, M., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19*(2), 313–330.

Mikheev, A. and L. Liubushkina (1995). Russian Morphology: An Engineering Approach. *Natural Language Engineering 3*(1), 235–260.

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. http://company.yandex.ru/articles/iseg-las-vegas.html.

Segalovich, I. and M. Maslov (1989). Dictionary-based Russian morphological analysis and synthesis with generation of morphological models of unknown words (in Russian). http://company.yandex.ru/articles/article1.html.

Segalovich, I. and V. Titov (2000). Automatic morphological annotation MYSTEM. http://corpora.narod.ru/article.html.

Véronis, J. (1996). MULTEXT-EAST (Copernicus 106). http://www.lpl.univaix.fr/projects/multext-east.

Wade, T. (1992). *A Comprehensive Russian Grammar*. Blackwell. 582 pp.

Yablonsky, S. A. (1999). Russian Morphological Analysis. In *Proceedings VEXTAL*.