Designing and Evaluating a Russian Tagset

Serge Sharoff[⋆], Mikhail Kopotev[∗], Tomaž Erjavec[†], Anna Feldman[‡], Dagmar Divjak[◇]

[⋆]University of Leeds,
Leeds, LS2 9JT, UK
s.sharoff@leeds.ac.uk
[∗]University of Helsinki,
P.O. Box 24, 00014, Helsinki, Finland
mihail.kopotev@helsinki.fi
[†]Jožef Stefan Institute,
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si
[‡]Montclair State University,
Montclair, NJ 07043, USA
feldmana@mail.montclair.edu
[◇]University of Sheffield
Sheffield, S10 2TN, UK
d.divjak@shef.ac.uk

Abstract

This paper reports the principles behind designing a tagset to cover Russian morphosyntactic phenomena, modifications of the core tagset, and its evaluation. The tagset and associated morphosyntactic specifications are based on the MULTEXT-East framework, while the decisions in designing it were aimed at achieving a balance between parameters important for linguists and the possibility to detect and disambiguate them automatically. The final tagset contains about 600 tags and achieves about 95% accuracy on the disambiguated portion of the Russian National Corpus. We have also produced a test set of tagging models and corpora that can be shared with other researchers.

1. Introduction

Historically, research on morphological analysis and disambiguation of Russian can be traced back to the very beginning of computational linguistics. The first programs of this sort were developed in 1950s in the context of machine translation, e.g., (Nikolaeva, 1958; Micklesen, 1958), see also (Hutchins, 1986). A milestone in this research was Zalizniak's Grammatical Dictionary of Russian, (Zalizniak, 1977), which provided a formal model of the diverse Russian morphology and led to a large number of implemented programs for Russian analysis and synthesis, e.g., (Mikheev and Liubushkina, 1995; Segalovich, 2003; Sokirko, 2004). These applications defined a set of rules for mapping between Russian forms and a set of formal morphological categories, e.g., number, gender, animacy.

However, these studies have not resulted in a tagset, i.e., a set of codes that combine the most important morphological categories describing each form into a single symbol. Some categories defined in Zalizniak (1977) are also less relevant for designing a tagset, e.g., impersonal verbs or pluralia tantum nouns, as they can lead to an unnecessary increase in the ambiguity.

Another problem is that Russian is a language with relatively free word order, and hence with a very rich morphology: it is morphology that plays a crucial role in signaling the syntactic relationships between the words in a sentence. This necessitates a rather extensive tagset. In addition, the relatively low number of morphemes forms, in particular those expressing case and number, yields a high level of ambiguity in individual forms. For example, the same form 'структуры' in different contexts can be interpreted in three different ways:

(1) анализ структуры
analysis structure$_{gen,sg}$
'analysis of the structure'

(2) в эти структуры
in these structure$_{acc,pl}$
'into these structures'

(3) эти структуры привлечены
these structure$_{nom,pl}$ involve$_{part,pass,perf,past,pl}$
'these structures are involved'

At the same time, morphological categories are defined on the level of individual words. The above-mentioned systems did implement the mapping between word forms and categories, but they did not have a disambiguation component. Hence, for a form they listed the complete set of options, e.g., $gen, sg; acc, pl; nom, pl$.

Stochastic taggers provide very efficient tools for automatic disambiguation in such cases, however, their performance has also not been studied for Russian. The problem here concerns the size of a tagset, which is typically much larger than for English, and consequent data sparsity, as a much larger corpus is needed to collect reasonable statistics if a tagset contains 500-2000 tags. Two exceptions are experiments done by Sokirko and Feldman with their colleagues (Sokirko and Toldova, 2005; Feldman et al., 2006), but their tagsets were not linguistically motivated: a Czech

tagset (Hajič and Hladká, 1998) was used in (Feldman et al., 2006), while the entire set of categories from (Sokirko, 2004) was used in (Sokirko and Toldova, 2005). Also their research has not produced publicly available tagging resources.

In this paper we present a linguistically motivated tagset for Russian, which is compatible with tagsets produced for other languages. We also evaluate its performance, and describe a publicly available resource that can be used in other experiments with tagging Russian.

2. Tagset principles

This section explains the MULTEXT-East (MTE) formalism as it is being developed for MULTEXT-East version 4 (Erjavec, 2006) and its application to Russian.

The MULTEXT-East language resources, currently available at version 3 (Erjavec, 2004) are a freely available multilingual dataset for language engineering research and development. The resources cover a large number of mainly Central and Eastern European languages, are standardised and their encoding harmonised across languages. They include the EAGLES-based morphosyntactic specifications, defining the features that describe word-level morphosyntactic annotations; medium scale morphosyntactic lexica; and annotated (parallel, comparable, speech) corpora. The tagsets defined by the morphosyntactic specifications have become a de-facto standard for many languages (e.g., Romanian, Croatian, Slovenian), and have also been used for languages developing morphosyntactic resources for the first time, e.g., Macedonian.

Currently in preparation is the fourth edition of the resources, which moves the encoding of the morphosyntactic specifications from LaTeX (plain text) to XML and adds several new languages, among them Russian.

The basic idea behind MULTEXT-type morphosyntactic specifications is to define, in a multilingual setting, main morphosyntactic categories (nouns, verbs, pronouns, ...) and their allowed attribute-value pairs, and to relate feature-structures describing morphosyntactic properties of words to compact strings, morphosyntactic descriptions (MSDs). For instance, the specifications can define that the feature structure

```
Noun, Type = common, Gender = masculine,
Number = singular, Case = accusative,
Animate = no
```

is valid, and that it corresponds to the MSD `Ncmsan`.

The new XML encoding of MULTEXT-type specifications allows for much greater flexibility in tagset design and use, such as shifting between language particular and common MTE tagsets, localisation of the tagset or feature set, etc. In this section we describe the application of these principles to Russian and the resulting resources, i.e., the specifications themsleves, and several derived formats, immediatelly useful for various mappings of the MDS.

2.1. The TEI specifications

The specifications are encoded in XML and follow the Text Encoding Initiative Guidelines (TEI) P5 (Sperberg-McQueen and Burnard, 2007). This allows for formal verification of their structure and offers a rich and well-documented XML element vocabulary, with supporting software.

The Russian specifications are encoded as a TEI document (with associated TEI P5 RelaxNG schema), which consists of the TEI header, giving the meta-data, introductory matter, a section for each defined category (PoS), and appendices.

The formal core of the specifications are tables with a constrained structure, one for each category, which define the attributes and their values and also give the positions of each attribute and one-letter codes for their values. The positions and codes allow for mapping between a feature-structure representation and its morphosyntactic description (MSD). An example of the encoding is given in Figure 1.

The structure of the tables is simple and largely self-explanatory in order to simplify their creation and maintenance. The specifications lend themselves well to localisation; in general, the accompanying text, type, attribute and value names and their codes can all be given also in other languages (say, `xml:lang="ru"`), allowing for interchangable encodings, even on the MSD level.

Another very important part of the specifications is the list of lexical MSDs, which should define all and only valid MSDs. Namely, while the feature-value tables define, to borrow XML terminology, well-formed MSDs, they do not constrain what combinations of feature-values are allowed, much less which MSD should exist for the language. The lexical MSDs are meant to give an exhaustive list of MSDs, optionally accompanying each by its expansion to a feature-structure, examples of usage and frequency information. For Russian, this information was taken from the disambiguated portion of the Russian National Corpus.

2.2. XSLT transforms

With XML as the base it becomes simple to offer various derived formats of the specifications by using XSLT stylesheets, also a W3C standard. The MTE-type specifications come with a XSLT library, offering the following outputs:

HTML As a reference, the specifications are best offered on the Web, in HTML. Producing it is done in two steps. First, the specifications — in particular the tables — are converted to a simplified TEI encoding and various indexes are generated, e.g., of categories, attributes and values. From this "print oriented" TEI format, standard TEI stylesheets are then used to arrive at the actual HTML.

TEI fsLib TEI is also suitable for encoding annotated corpora, and allows for linking corpus tags to their definitions, themselves encoded as feature-structures. On the basis of the lexical list the XSLT produces a feature-structure library, which defines the MSD ID, and gives its decomposition to attribute-value pairs, and, on the basis of the tables, produces a feature library, giving the names of the attributes and values.

Tabular expansions Here, on the basis of the specifications, a list of MSDs is expanded or translated into

```
<div>
  <head xml:lang="en">Attribute-value table</head>
  <table n="msd.cat" xml:id="msd.cat.N-ru" select="ru">
    <head xml:lang="en">Specification for Noun</head>
    <row role="type">
      <cell role="position">0</cell>
      <cell role="name" xml:lang="en">Noun</cell>
      <cell role="code" xml:lang="en">N</cell>
    </row>
    <row role="attribute">
      <cell role="position">1</cell>
      <cell role="name" xml:lang="en">Type</cell>
      <cell role="values">
        <table>
          <row role="value">
            <cell role="name" xml:lang="en">common</cell>
            <cell role="code" xml:lang="en">c</cell>
          </row>
          <row role="value">
            <cell role="name" xml:lang="en">proper</cell>
            <cell role="code" xml:lang="en">p</cell>
          </row>
        </table>
      </cell>
    </row>
    ...
```

Figure 1: The MULTEXT-East Russian morphosyntactic specifications: start of Noun.

various formats, e.g., a short and long expansion into attribute-value pairs, collating sequence, localisation into another language (if given in the specifications), various normal forms, etc. If the XSLT output method is set to XML, the result is given as TEI tables; if to text, the result is a tab separated file.

2.3. Available Resources

The MULTEXT-East morphosyntactic specifications and MSD tagset for Russian are being made available in the source XML/TEI and several derived formats:

- HTML format, suitable for browsing and reading.
- List of lexical MSDs in tabular format, with the following structure, by column:
  - MSD: the morphosyntactic description;
  - collation string: sorting MSDs by this string will give them in the order in which features are defined in the tables. This is useful for presenting MSDs in lists, as languages have traditional orderings of attribute values (e.g., nominative, genitive, dative,...);
  - long expansion: a decomposition of the MSD into category and list of attribute=value pairs;
  - short expansion: a decomposition of the MSD into values only, The short expansion is thus similar to just retaining values from the long expansion, but it additionaly decorates certain values, e.g., a Animate=no is displayed as -Animate.

  This format is useful for various manipulations and further explotation of MSDs, such as sorting, processing of the MSDs as a feature-structures, or giving a synoptic description of the MSDs; e.g., a Web concordancer displaying MSDs might return Afp as `<span title="Adjective qualificative positive">Afp</span>`
- List of lexical MSD in tabular format in positional notation:
  - first column contains the MSD
  - second column contains its category
  - each of the following columns contains the value of one attribute, regardless of the category.

  This transformation of the MSDs is useful for enabling access to individual attributes, e.g., for making queries to corpora via the Corpus Workbench (`http://cwb.sf.net`). For instance, to return all tokens marked as feminine genitiv regardless of the part-of-speech the CQP query would be: `[gender="feminine" & case="genitive"]`

3. Properties of the tagset

The design of the Russian tagset proceeded in several steps. First, taking into account the linguistic structure of the Russian language, as well as features marked in current large corpora of Russian, the appropriate features were selected or added to the MTE specifications. Then, a procedure was written that maps from representations used in several corpora to the MTE tagset, with the resulting MSDs validated against the specifications.

In our choices as to what attributes / values to define for Russian we tried to take into account:

- the balance between parameters important for linguists and the possibility of their automatic detection;

- the availability of features in existing corpora that can be used for training;
- the possibility to disambiguate features using local context;
- the possibility to share the tagset with other Slavonic languages to create, in perspective, a common Slavonic morphological tagset.

Two Russian morphologically annotated corpora have been taken into account in designing the tagset:

- the disambiguated portion of the Russian National Corpus (Sharoff, 2005), which is comparable to the BNC Sampler in its size and accuracy of annotation, and
- HANCO, developed in the University of Helsinki (Kopotev and Mustajoki, 2003)

There are some other Russian corpora, however the chosen ones represent two extremes in Russian morphological annotation. The RNC tagset is built in order to simplify the process of annotation, while the HANCO is more accurate, because it is directed toward a wider circle of end-users, including L2 teachers and learners of Russian. Nevertheless, the actual difference in the number of categories in the two corpora is small, 137 labels used in the RNC against 147 in the HANCO.

The resulting tagset we developed for Russian contains the 12 MTE main categories: noun, verb, adjective, pronoun, adverb, adposition, conjunction, numeral, particle, interjection, abbreviation, and residual (the last one is reserved for special purposes). They have 0-10 attributes each, in total giving 156 attribute-value pairs, which overlap with categories used in the tagsets of the above-mentioned corpora. Some particular decisions about what is in the tagset and what is not, have been made on the basis of some more or less general strategies.

For instance, multiword expressions have been avoided as much as possible, even if they have been marked in the corpora, as their use in the tagset complicates processing, for argumentation see (Kilgarriff, 1997). At the same time, analytic forms, like буду летать ('will fly'), are represented using the `Type=Auxilliary` for the first lemma.

Likewise some infrequently used homonymic forms have been excluded, for instance, the paucal case of nouns (три шагá '[three] steps$_{pauc}$) because it is fully homonymic to the genitive case in all written contexts (however, they differ in an accent place). Similarly, impersonal verbs (those, having no Subject argument светать 'dawn'), as their inflexion is fully homonymic to third person, singular.

We did not include pluralia tantum of nouns, e.g., очки, 'spectactles', as this is a lexical feature that does not change the syntactic function of the word. Many pluralia tantum nouns are also homonymic to the plural forms of ordinary nouns (очки, 'points').

The MTE specifications for Russian also use some features that go beyond traditional grammars, e.g., (Zalizniak, 1977). The reasons that were behind their addition are:

- they can be detected automatically using local context;
- they exist in some other Slavonic languages;
- they are marked at least in one of the existing corpora.

For instance, the vocative case (Пап!, 'Dad$_{voc}$') is rarely recognised in traditional grammars. However, it exists in Polish, Ukrainian etc. and, thus, should be included into the Slavonic tagset. This case has been marked in the RNC as well.

In Slavonic languages, aspect is a verbal category, which is obligatorily marked on all verbal forms. However, for certain Russian verbs it is difficult to define which aspect they instantiate as the same form is used in both imperfective and perfective contexts, e.g., дешифровать, 'decode'. These verbs are annotated as biaspectual in HANCO, and that tag is preserved in the tagset presented in this paper.

Special attributes in the tagset are used to mark two non-standard cases to nouns, Partititve and Locative$_2$. Morphologically, these cases have clear distinguishing features in both inflection and semantics. However, both govern adjective forms in Genitive and Locative respectively, e.g., partitive налить теплого чаю 'to pour hot$_{gen}$ tea$_{part}$', and locative в зеленом лесу, in a green$_{loc}$ forest$_{loc2}$.

Thus, morphologically they are individual cases, while syntactically they are indistinguishable from, respectively, Genitive and Locative. Within the tagset, such nouns will still be marked in the Genitive or Locative case, and information about their morphological features can be preserved in the Case2 attributes (the values are Partitive and Locative$_2$).

We have also designed means for mapping from corpus-specific representations of morphological features to MSD. For instance, the disambiguated portion of the RNC uses an HTML-like representation to store morphological information

```
<span lex='структура'
class='S=f,gen,inan,sg'>структуры</span>
```

which can be converted to the MSD representation as:

структуры `Ncfsgn` структура

4. Evaluation and Discussion

The problem of tagset design is particularly important for highly inflected languages. If all of the syntactic variations that are realized in the inflectional system were represented in the tagset, there would be a huge number of tags, and it would be practically impossible to implement or train a simple tagger.

Elworthy (1995) defines the external and internal criteria for designing a tagset. The external criterion is that the tagset must be capable of making the linguistic (for example, syntactic or morphological) distinctions required in the output corpora. The internal criterion on tagsets is the design criterion of making the tagging as accurate as possible. It can be argued that a smaller tagset should improve the accuracy of tagging, since it puts less of a burden on the tagger to make fine distinctions. In information-theoretic terms, the number of decisions required is smaller, and hence the tagger needs to contribute less information to make the decisions. A smaller tagset may also mean that more words have only one possible tag and so can be handled trivially. On the other hand, more detail in the tagset may help the tagger. For example, if nouns and adjectives that modify these nouns are marked for case, gender, and number, the tagger can effectively model agreement in simple noun

Accuracy	Overall	Known W	Unknown W
TnT	95.28%	96.27%	66.64%
TT	93.50%	94.33%	62.44%
SVMTool	92.24%	93.26%	54.28%

Table 1: Overall accuracy of TnT, TT, and SVMTool, full tagset

phrases, by having a higher probability for a singular nominative feminine adjective followed by a singular nominative feminine noun than it does for a singular nominative feminine adjective followed by a plural genitive masculine noun.

Our question, both theoretical and practical, is what tagset design best serves Russian. We evaluate the tagset both qualitatively and quantitatively.

4.1. Quantitative evaluation

We started with the disambiguated portion of the Russian National Corpus (about 5 million words), converted it into the MSD representation, and trained three statistical POS taggers: TnT (Brants, 2000), TreeTagger (Schmid, 1994) and SVM Tagger (Giménez and Màrquez, 2004). 10% of the corpus was held out for testing the performance of the three taggers.

The overall accuracy is given in Table 1. The TnT tagger reaches the state-of-the-art performance with our corpus and tagset. TreeTagger (TT) is slightly behind. In spite of the promising accuracy on small tagsets (e.g., Penn Treebank), on our tagset SVM Tool happened to be the slowest to train (one iteration takes about 24 hours vs. less than one minute by either TT and TnT), it showed the slowest tagging on the test set (40 min vs. 8-10 sec) and was the least accurate. More research is needed to determine if SVM-based tagging is applicable to large tagsets.

Among the three taggers, TnT was the best, so we decided to look at its performance on individual positions of the major open classes: nouns, verbs, and adjectives (see Table 2, Table 3, and Table 4, respectively). We also measured the performance of TnT on a reduced tagset of Russian, which is roughly comparable to the Penn Treebank tagset (Marcus et al., 1993). The performance of TnT reached 97.09% on the reduced Russian tagset, which is only 2% better than the performance of the tagger on the full detailed tagset we developed in this project.

The evaluation on individual categories reveals that the most difficult part of speech category is the category of nominals, which includes adjectives and nouns. The results of evaluation show that gender and case are the most challenging attributes. One of the plausible explanations that we can offer here is that even though case and gender participate in syntactic agreement in Russian, these categories are more idiosyncratic than, say, person or tense. Further work should include an attempt to quantify the extent to which gender and case difficulties are due to pure lexical idiosyncrasy.

	known	unknown
full tag	90.99	56.05
1: category	99.02	93.61
2: type	98.42	86.00
3: gender	97.51	77.23
4: number	97.89	89.26
5: case	93.03	80.23

Table 2: Accuracy of TnT on nouns in % (full tagset).

	known	unknown
full tag	96.34	73.12
1: category	99.00	93.74
2: type	99.00	93.74
3: vform	98.61	91.44
4: tense	97.69	84.10
5: person	98.93	93.33
6: number	98.80	93.42
7: gender	98.95	93.57
8: voice	98.89	93.01
9: definiteness	98.97	93.60
10: aspect	96.93	75.23
11: case	98.98	93.68

Table 3: Accuracy of TnT on verbs in % (full tagset).

We also varied the size of the training corpus to see how it affects the performance of TnT with our tagset. What seems to come out of our experiments is that there is a consistent relationship between the size of the tagset and the tagging accuracy for Russian.

The tagset we developed uses the positional tag system. There are several reasons why this scheme is preferred over the unordered compact tag. First, the morphological descriptions are more systematic. In each system, the attribute positions are determined by the Category. Thus, for example, knowing that a token is a noun (N) automatically provides information that the gender, number, and case positions should have values. This kind of description seems to be the most adequate for languages with rich morphology. Second, the fact that a tag can be decomposed into individual components has been used in various applications, e.g., (Hladká, 2000; Hana et al., 2004). Moreover, the evaluation can be done in a more systematic way. Each category can be evaluated separately on each morphological feature. Not only is it easy to see on which POS the tagger performs the best/worst, but it is also possible to determine which individual morphological features cause the most problems. Finally, the translation between various positional tagsets is easy, since the formalism states clearly what attribute occupies what position in the tag.

4.2. Qualitative evaluation

A manual spotcheck of the test corpus tagged by TT revealed several recurrent errors; all of these concern forms

	known	unknown
full tag	89.13	80.51
Tag slot		
1: category	97.25	91.72
2: type	97.25	91.72
3: degree	97.24	91.72
4: gender	95.67	89.77
5: number	97.00	90.98
6: case	90.54	84.37

Table 4: Accuracy of TnT on adjectives in % (full tagset).

that are shared across cases. Typically, the accusative singular takes priority over the nominative singular for inanimate masculine nouns and for neuter nouns as well as for feminine nouns with the soft sign (мягкий знак); the same happens in the plural for feminine nouns. Also, the genitive singular takes priority over the nominative/accusative plural for feminine nouns; something similar is attested for the accusative singular of animate masculine nouns. Thirdly, the genitive singular takes priority over the instrumental for feminine adjectives and sometimes an instrumental is used instead of a genitive for feminine singular adjectives, or vice versa.

Interestingly, an experiment with ten second year British students of Russian revealed that students are not able to spot the errors produced by the TT tagger: intermediate level students too seem to analyze forms in isolation.

TnT performs considerably better, producing only a fraction of the errors outputted by TT. Only two mistakes are encountered: the genitive singular takes priority over the nominative/accusative plural for feminine nouns, and occasionally accusative singular is selected instead of nominative singular for inanimate masculine nouns, for neuter nouns and for feminine nouns in a soft sign.

5. Conclusions and future work

This research resulted in a tagset and tagging resources for Russian. The tagset documentation and models for TnT, TT and SVM Tagger can be downloaded from `http://corpus.leeds.ac.uk/mocky/`.

However, RNC and Hanco corpora used in this study cannot be made freely available because of copyright restrictions. Therefore, the lack of annotated corpora to train taggers can affect further research into tagging Russian and developing new resources by other groups.

To respond to this need, we selected a subset of the Russian Internet corpus (Sharoff, 2006), made sure that it represents a variety of text types (fiction, news, discussion forums, research papers and tutorials) and tagged it with the three tools we tested. The corpus consists of 843 texts and contains about 5 mln words. As shown in the previous section the accuracy of tagging is not 100%, so the corpus cannot be compared to manually tagged resources. However, it is available to other researchers, while its accuracy can be improved by comparing the errors made by different taggers on it using majority voting. It already contains three tagged versions, while if any other tagger for Russian becomes available, the resource can be improved further.

We think that our resources will become the basis for developing new Russian language processing tools. Morphological tags carry important information, which is essential for syntactic parsing or text-to-speech applications, for instance. In Russian, it is often not enough to distinguish verbs from nouns only — we need more detailed morphosyntactic descriptions of lexical items. For example, in order to determine which syllable of a given instance of the Russian word снега should be stressed, one must know the morphological properties of that instance – the genitive singular form of the word is stressed on the first syllable, while the nominative plural form is stressed on the second: сне́га (snow$_{gen.masc.sg}$) vs. снега́ (snow$_{nom-acc.pl}$).

We also think that the standard morphosyntactic specifications we developed for Russian make it possible to create a specification for other Slavonic languages that are less commonly studied in corpus linguistics, such as Ukrainian and Belorussian, both closely-related to Russian. The tagset presented in this paper will also make it easier to create multilingual applications or to evaluate language technology across several languages, even if they are not closely related, such as English or Persian, since their morphology is also described in the MTE tagset. Finally, a shared tagset provides a basis for various studies in language typology.

6. References

Brants, Thorsten. 2000. TnT - a statistical part-of-speech tagger. In *Proc. of 6th Applied Natural Language Processing Conference*, pages 224–231, Seattle.

Elworthy, David. 1995. Tagset design and inflected languages. In *7th Conference of the European Chapter of the Association for Computational Linguistics (EACL), From Texts to Tags: Issues in Multilingual Language Analysis SIGDAT Workshop*, pages 1–10, Dublin.

Erjavec, Tomaž. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, pages 1535 – 1538.

Erjavec, Tomaž. 2006. Multext-east morphosyntactic specifications and xml. In Slavcheva, Milena, Simov, Kiril, and Angelova, Galia, editors, *Readings in multilinguality*, pages 41–48. Bulgarian Academy of Sciences, Sofia.

Feldman, Anna, Hana, Jirka, and Brew, Chris. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proc. LREC 2006*.

Giménez, Jesús and Màrquez, Lluís. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the Forth Language Resources and Evaluation Conference, LREC 2004*, Lisbon.

Hajič, Jan and Hladká, Barbora. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of COLING-ACL*, pages 483–490.

Hana, Jirka, Feldman, Anna, and Brew, Chris. 2004. A Resource-light approach to Russian morphology: tagging Russian using Czech resources. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 222–229, Barcelona, Spain.

Hladká, Barbora. 2000. *Czech Language Tagging*. Ph.D. thesis, Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics (MFF), Charles University (UK), Prague, Czech Republic.

Hutchins, W. J. 1986. *Machine translation: past, present, future*. John Wiley & Sons.

Kilgarriff, Adam. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.

Kopotev, Mikhail and Mustajoki, Arto. 2003. Principy sozdanija Hel'sinkskogo annotirovannogo korpusa russkih tekstov (HANCO) v seti internet. *Naučno-tehničeskaja informacija, Ser. 2*, (5):33–37. In Russian.

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

Micklesen, L.W. 1958. Russian-English MT. In *American contributions to the Fourth International Congress of Slavicists*, Moscow.

Mikheev, Andrei and Liubushkina, Liubov. 1995. Russian morphology: An engineering approach. *Natural Language Engineering*, 1(3):235–260.

Nikolaeva, T.N. 1958. Soviet developments in machine translation: Russian sentence analysis. *Mechanical Translation*, 5(2):51–59.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester.

Segalovich, Ilya. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *Proc. of MLMTA-2003*, Las Vegas.

Sharoff, Serge. 2005. Methods and tools for development of the Russian Reference Corpus. In Archer, D., Wilson, A., and Rayson, P., editors, *Corpus Linguistics Around the World*, pages 167–180. Rodopi, Amsterdam.

Sharoff, Serge. 2006. Creating general-purpose corpora using automated search engine queries. In Baroni, Marco and Bernardini, Silvia, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. http://wackybook.sslmit.unibo.it.

Sokirko, Alexei and Toldova, Svetlana. 2005. Sravnenie effektivnosti dvuh metodik snyatiya lexicheskoy i morfologicheskoy neodnoznachnosti dlya russkogo yazyka. Technical report, `http://www.aot.ru/docs/RusCorporaHMM.htm`. In Russian.

Sokirko, A.V. 2004. Morphologicheskie moduli na sajte `www.aot.ru`. In *Proc. DIALOG'04*. In Russian.

Sperberg-McQueen, C. Michael and Burnard, Lou, editors. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford, 5 edition.

Zalizniak, A.A. 1977. *Grammaticheskiyj Slovar' Russkogo Jazyka*. Russkij Jazyk.