## POS tagging for English

<table>
<thead>
<tr>
<th>System name</th>
<th>Short description</th>
<th>Main publication</th>
<th>Software</th>
<th>Extra Data?</th>
<th>All tokens</th>
<th>Unknown words</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>TnT*</td>
<td>Hidden markov model</td>
<td>Brants (2000)</td>
<td>TnTIP</td>
<td>No</td>
<td>98.46%</td>
<td>65.86%</td>
<td>Academic/research use only (license required)</td>
</tr>
<tr>
<td>MEMM</td>
<td>MEMM with external lexical information</td>
<td>Denis and Sagot (2006)</td>
<td>Apage linguistic workbench</td>
<td>No</td>
<td>96.66%</td>
<td>91.29%</td>
<td>CeCILL-C</td>
</tr>
<tr>
<td>GENIA Tagger*</td>
<td>Maximum entropy cyclic dependency network</td>
<td>Tsuruoka, et al (2006)</td>
<td>GENIAIP</td>
<td>No</td>
<td>97.05%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>Averaged Perceptron</td>
<td>Averaged Perception discriminative sequence model</td>
<td>Collins (2002)</td>
<td>Not available</td>
<td>No</td>
<td>97.11%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>Maxent-easiest-first</td>
<td>Maximum entropy bidirectional easiest-first inference</td>
<td>Tsuruoka and Tsujii (2005)</td>
<td>Easiest-firstIP</td>
<td>No</td>
<td>97.15%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>SVMTool</td>
<td>SVM-based tagger and tagger generator</td>
<td>Giménez and Marquez (2004)</td>
<td>SVMToolIP</td>
<td>No</td>
<td>97.16%</td>
<td>89.01%</td>
<td>LGPL 2.1</td>
</tr>
<tr>
<td>LAPPOS</td>
<td>Perception-based training with lookahead</td>
<td>Tsuruoka, Miyao, and Kazama (2011)</td>
<td>LAPPOSIP</td>
<td>No</td>
<td>97.22%</td>
<td>Not available</td>
<td>MIT</td>
</tr>
<tr>
<td>Morfeu/COMPOST</td>
<td>Averaged Perceptron</td>
<td>Spoustová et al. (2009)</td>
<td>COMPOSTIP</td>
<td>No</td>
<td>97.23%</td>
<td>Not available</td>
<td>Non-free (academic-only)</td>
</tr>
<tr>
<td>Morfeu/COMPOST</td>
<td>Averaged Perception</td>
<td>Spoustová et al. (2009)</td>
<td>COMPOSTIP</td>
<td>Yes</td>
<td>97.44%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>Stanford Tagger 1.0</td>
<td>Maximum entropy cyclic dependency network</td>
<td>Toutanova et al. (2003)</td>
<td>Stanford TaggerIP</td>
<td>No</td>
<td>97.24%</td>
<td>89.04%</td>
<td>GPL v2+</td>
</tr>
<tr>
<td>Stanford Tagger 2.0</td>
<td>Maximum entropy cyclic dependency network</td>
<td>Manning (2011)</td>
<td>Stanford TaggerIP</td>
<td>No</td>
<td>97.29%</td>
<td>89.70%</td>
<td>GPL v2+</td>
</tr>
<tr>
<td>Stanford Tagger 2.0</td>
<td>Maximum entropy cyclic dependency network</td>
<td>Manning (2011)</td>
<td>Stanford TaggerIP</td>
<td>Yes</td>
<td>97.32%</td>
<td>90.79%</td>
<td>GPL v2+</td>
</tr>
<tr>
<td>LTAG-spinal</td>
<td>Bidirectional perceptron learning</td>
<td>Shen et al. (2007)</td>
<td>LTAG-spinalIP</td>
<td>No</td>
<td>97.33%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>SCCN</td>
<td>Semi-supervised condensed nearest neighbor</td>
<td>Sogaard (2011)</td>
<td>SCCNIP</td>
<td>Yes</td>
<td>97.50%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>ChartWNN</td>
<td>MLP with Neural Character Embeddings</td>
<td>dos Santos and Zaenen (2014)</td>
<td>Not available</td>
<td>No</td>
<td>97.32%</td>
<td>89.86%</td>
<td>Unknown</td>
</tr>
<tr>
<td>structReg</td>
<td>CRFs with structure regularization</td>
<td>Sun (2014)</td>
<td>Not available</td>
<td>No</td>
<td>97.36%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>BLSTM-CRF</td>
<td>Bidirectional LSTM-CRF Model</td>
<td>Huang et al. (2015)</td>
<td>Not available</td>
<td>No</td>
<td>97.55%</td>
<td>Not available</td>
<td>Unknown</td>
</tr>
<tr>
<td>NLPI4J</td>
<td>Dynamic Feature Induction</td>
<td>Choi (2016)</td>
<td>NLPI4JIP</td>
<td>Yes</td>
<td>97.64%</td>
<td>92.03%</td>
<td>Apache 2</td>
</tr>
</tbody>
</table>
POS tagging for English and Russian

- **POS-tagging for English:**
  - Relatively simple morphology.
  - Established training corpus (WSJ Penn Treebank).
  - Multiple approaches (HMM, CRF, dependency networks, neural network methods).
  - High baseline.

- **POS-tagging for Russian:**
  - Large number of tags.
  - Ubiquitous homonymy.
  - Long-distance dependencies.
  - Fine-grained categories, complex interaction between them.
POS tagging for Russian

- POS-tagging algorithm for Russian:
  - No reference corpora.
  - No comparison of different algorithms.
  - Problems with baseline approaches.

- Possible algorithms:
  - HMM cannot extract all the information.
  - CRF require too much memory, cannot handle long distance dependencies.
  - Discriminative-based approaches not tested, possible large number of features.
  - Neural network approaches not tested.

- Our goals:
  - Provide a reference corpus.
  - Compare different algorithms.
  - Determine directions for future work.
Previous Russian Morphology forum:

Ru-Eval 2010 * state-of-the-art, mostly rule-based taggers, test dataset (2k words)

- POS & Lemmatization: 13 answers
- Morphology: 12 answers
- Rare words: 8 answers

- Disambiguation (POS, Lemma): 7 answers

Soft evaluation: 94.5 - 95% accuracy
1. **Closed track**: the participants are allowed to train their models only on provided data.
   - for research groups and student teams
   - own dictionaries allowed

2. **Open track**: track members are allowed to bring any data for learning
   - for enterprise participants

Full morphological tags are evaluated and also (optionally) lemmatization.
Universal Dependencies 1.4 and 2.0

Parts of Speech:
noun (NOUN), proper name (PROP), adjective (ADJ), pronoun (PRON) numeral (NUM), verb (including auxiliary, VERB),
adverb (ADV), determinant (DET), conjunction (CONJ), preposition (ADP), particle (PART), interjection (INTJ).

Also marked: punctuation marks (PUNCT), non-word tokens (X), parenthesys (H).

Omitted: SYM (symbol) and AUX (auxiliary verb).
## Tagset

<table>
<thead>
<tr>
<th>Case</th>
<th>nominative - Nom, genitive - Gen, dative - Dat, accusative - Acc, locative – Loc, instrumental - Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>masculine - Masc, feminine - Fem, neuter - Neut</td>
</tr>
<tr>
<td>Number</td>
<td>singular - Sing, plural - Plur</td>
</tr>
<tr>
<td>Animacy</td>
<td>animate - Anim, inanimate - Inan</td>
</tr>
<tr>
<td>Tense</td>
<td>past - Past, present or future - Notpast</td>
</tr>
<tr>
<td>Person</td>
<td>first – 1, second – 2, third - 3</td>
</tr>
<tr>
<td>VerbForm</td>
<td>infinitive - Inf, finite - Fin, gerund - Conv (participles are treated as ADJ)</td>
</tr>
<tr>
<td>Mood</td>
<td>indicative – Ind, imperative - Imp</td>
</tr>
<tr>
<td>Variant</td>
<td>short form – Brev (no tag for long form)</td>
</tr>
<tr>
<td>Degree</td>
<td>positive or superlative - Pos, comparable - Cmp</td>
</tr>
<tr>
<td>NumForm</td>
<td>numeric token – Digit (if the token is written in alphabetic form, no mark is placed).</td>
</tr>
</tbody>
</table>
Tagset problems and solutions

Categorical mismatches in different data sources:

- no Aspect tags on GICR & Syntagrus data: Tense (past, present, future) → Tense (past, notpast)
- PROPN and NOUN are equally evaluated
- There is a special “H” tag for parenthetic constructions
- Participles and ordinal numerals are considered adjectives, gerunds - part of a verb paradigm
- Predicatives are considered short forms of adjectives, with an exception for “нет”, which is a verb

Some categories listed explicitly: Determiner (UD), Conjunctions, Particles, Prepositions, Parenthesis, Pronouns.

Conj, Part, Prep, Int, H, X and some adverbs (как, пока, так, когда), homonymic to them, are not taken into account during evaluation. GICR data was proposed as a standard if any other differences occurred.
For both tracks we provide the following training data:

annotated data:

1) RNC Open: a manually disambiguated subcorpus of the Russian National Corpus - 1.2 million words (fiction, news, nonfiction, spoken, blog)
2) GICR corpus with the resolved homonymy - 1 million words
3) OpenCorpora.org data - 0.4 million tokens
4) UD SynTagRus - 0.9 million tokens (fiction, news)

And also plain text data: 1) LiveJournal (from GICR) 30 million words 2) Facebook, Twitter, VKontakte - 30 million words 3) Librusec - 300 million words

Test set:
1. news texts (Lenta.ru)
2. fiction (Russian Magazine Hall, magazines.russ.ru)
3. social networks (vk.com) – from unpublished part of GICR materials for MorphoRuEval (other data resources were previously published).

600-900 thousand tokens for each segment.

Gold standard
3 different segments from GICR for testing – Lenta.ru, fictions (Russian Magazine Hall), VK
7000 tokens each

Baseline:
TreeTagger trained on annotated MorphoRuEval data
best - 79% accuracy per tag depending on the sources of testing and training data
best - 26% accuracy per sentence.
Evaluation and metrics

Four metrics for ranking:

- Percentage of correct tags and tag-lemma pairs (in case the system outputs lemmas).
- Percentage of correctly labeled sentences both by tags and by tag-lemma pairs.
- All metrics are calculated for three subtasks and for the whole dataset.
- Sentence accuracy on the entire dataset used for ranking.

Reason: subtle differences in tag accuracy become significant for sentence accuracy.

We also used the following conventions:

1) Both PROPN and NOUN labels for proper nouns is correct. The same holds for SCONJ and CONJ with respect to conjunctions.
2) Capitalization is not significant for lemmatization.
3) è and ę are not distinguished.
Evaluated categories

Evaluated POS and grammemes:

- Nouns (gender, number, case).
- Adjectives (gender, number, case, degree, brevity).
- Verbs (mood, tense, person, gender, number).
- Determiners (gender, number, case).
- Pronouns (gender, number, case, person).
- Numerals (gender, case, graphic form)
- Adverbs (degree)

Not evaluated:

- Conjunctions,
- Prepositions,
- Particles,
- Parentheses,
- Punctuation.

Other information provided by participants is not taken into account
## Competition results

### Top 6 teams (of 15 participants)

<table>
<thead>
<tr>
<th>Team name</th>
<th>Track</th>
<th>Tags</th>
<th>Sents</th>
<th>Lemma</th>
<th>Lemma sents</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABBYY</td>
<td>Open</td>
<td>97.11</td>
<td>83.68</td>
<td>96.91</td>
<td>82.13</td>
</tr>
<tr>
<td>MSU-1</td>
<td>Closed</td>
<td>93.39</td>
<td>65.29</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IQMEN</td>
<td>Closed</td>
<td>93.08</td>
<td>62.71</td>
<td>92.22</td>
<td>58.21</td>
</tr>
<tr>
<td>Sagteam</td>
<td>Closed</td>
<td>92.64</td>
<td>58.4</td>
<td>80.73</td>
<td>25.01</td>
</tr>
<tr>
<td>Aspect</td>
<td>Closed</td>
<td>92.57</td>
<td>61.01</td>
<td>91.81</td>
<td>56.49</td>
</tr>
<tr>
<td>Morphobabushka</td>
<td>Closed</td>
<td>90.07</td>
<td>48.1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Several types of algorithms for morpho tagging:

- Neural networks (ABBYY (clear winner), Sagteam, Aspect)
- Classification-based (IQMEN, Morphobabushka).
- Reranking-based (MSU).

Lemmatization algorithms: usually a conversion pattern is guessed using the same features as for the tag itself.

Most of the participants train on GICR subset of the training data.

ABBYY team additionally train on Wiktionary corpus annotated by Compreno parser.
Top-ranked algorithms

- The clear winner **ABBYY team**
- LSTM network as main classifier.
- Several layers in the network (up to 10).
- Two types of features on input layer:
  - Grammatical and suffix features extracted using Compreno parser.
  - Pre-trained word embeddings fine-tuned on the training set.
Second team: MSU, winner on the closed track. An attempt to build linguistically oriented system.

HMM as basic classifier generating hypotheses.

Initial training data extended with transitivity for verbs and case for prepositions.

Hypotheses are reranked using high-level features.

Examples of features:

- Number of coordinated adjective-noun groups.
- Number of coordinated preposition-noun groups.
- Number of nominative nouns coordinated with verbs.
- Number of transitive verbs having a direct object.

Learning algorithm: generate hypotheses for the sentences in the training set and train a linear classifier on the differences between top hypothesis and the others.

Decision algorithm: select the hypothesis with the highest score according to the classifier.
Top-ranked algorithms

- **Aspect team**: bidirectional LSTM.
  - Separate character, flexion and stem embedding on the input layer.
  - Embeddings for stem and flexions trained on LibRuSec corpora.
  - Several dense layers in the network.

- **Sag**: convolutional neural network.
  - Character level embedding on the input layer for individual words.
  - Several layers in the network.
Top-ranked algorithms

- **IQMEN team**: window-based classification approach.
  - Each word is tagged in isolation using the features from surrounding words and the word itself.
  - Word features: word prex and sux up to 4 characters.
  - Left context features: POS and tag features (case, gender, number) for all the words in 7-word window.
  - Right context features: ambiguity classes for POS and tag features.
  - SVM with hash kernel as the final classifier.

- **Morphobabushka team**:
  - No dictionary used, word is tagged using only features.
  - Word features: character ngrams. Features are extracted from the word itself and its neighbours.
  - A classifier outputs the labels for tag features (e.g. case) or combinations of features (number + gender).

- **NB-SVM classier**
Questions for future work

- Main problems for tagging morphologically rich languages:
  - Abundancy of morphological features and classes.
  - Long distance dependencies.
- Ways to overcome feature abundancy:
  - More powerful classifiers (neural networks).
  - Modification of more traditional classifiers in window-based approaches.
- Ways to deal with long-distance dependencies:
  - Convolutional layers in CNN, long-short memory in LSTM.
  - Global features in reranking.
  - Global features in classification (not used?).
- Role of training corpora and dictionaries?
  - Character-based approaches allow to work without dictionaries.
  - Neural networks require large corpora (though unannotated).
- Global approaches on the top of neural network model (neural network provides local correctness, global features stand for global constraints).
Comparing to previous evaluation of morphological parsers for Russian language, current systems show significant improvement. Indeed, the top-ranked of the [Lyashevskaya et al., 2010] competition achieved 97% result only for POS-tagging, while the winner of current competition showed the same result for entire grammatical tags. The top-system result is comparable with results for other inflective languages with free word order and rich inflective morphology, such (95.75% for Czech in [Strakova, 2013]).

Shared task on morphological tagging showed fruitful results in several important aspects:

- An original data set collected from different corpora which was annotated in a single format consistent with UD guidelines was prepared and presented;
- The comprehensive guidelines for testing procedure and evaluation were created.
- The comparison of different parsing strategies showed that neural network approach is state-of-the-art method for morphological parsing of Russian.
- A dataset for future improvement of morphological parsers, comprising texts from different sources, was created.