Designing and Evaluating a Russian Tagset

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

[⋆]University of Leeds, [*]University of Helsinki,[†]Jožef Stefan Institute, [‡]Montclair State University,[◇]University of Sheffield

s.sharoff@leeds.ac.uk, mihail.kopotev@helsinki.fi, tomaz.erjavec@ijs.si, feldmana@mail.montclair.edu, d.divjak@shef.ac.uk

1. Existing research

- MT in 1950s – hand-crafted ad-hoc rules
- Zalizniak (1977) – formalisation of Russian morphology
- Mikheev, Segalovich, Sokirko, Starostin – various implementations of Russian analysis and synthesis
- Sokirko, Feldman – experiments with statistical taggers, but no resources

2. Problems with Russian

Free word order and very rich morphology it is morphology that plays a crucial role in signaling the syntactic relationships.

Low number of morpheme forms The same form in different contexts can be interpreted in 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 к
these structure$_{nom,pl}$ involve$_{part,pass,perf,past,pl}$ to
'these structures are involved in...'

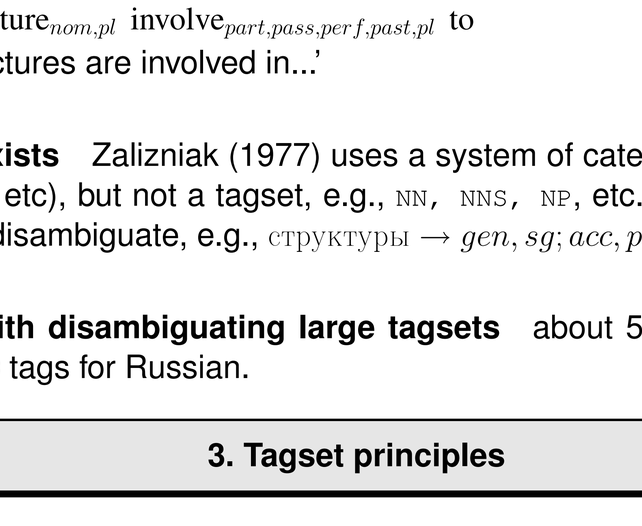
No tagset exists Zalizniak (1977) uses a system of categories (case, gender, number, etc), but not a tagset, e.g., `NN`, `NNS`, `NP`, etc. Analysis applications do not disambiguate, e.g., структуры $\rightarrow gen, sg; acc, pl; nom, pl$.

Problems with disambiguating large tagsets about 50 tags for English vs. 500-2000 tags for Russian.

3. Tagset principles

MULTEXT-East (MTE) A freely available multilingual dataset for language engineering research and development. The resources cover a large number of mainly Central and Eastern European languages.

MTE morphosyntactic specifications Main morphosyntactic categories (nouns, verbs, pronouns, ...) and their allowed attribute-value pairs. Feature-structures describing morphosyntactic properties of words can be mapped to compact strings, morphosyntactic descriptions (MSDs).



`<span lex='структура' class='S=f,gen,inan,sg'>`структуры`</span>`
→ структуры `Ncfsgn` структура
→ `Noun, Type = common, Gender = feminine, Number = singular, Case = genitive, Animate = no`

4. Properties of the tagset

- the balance between parameters important for linguists and the possibility of their automatical detection;
- the availability of features in existing corpora that can be used for training;
- the possibility to share the tagset with other Slavonic languages to create, in perspective, a common Slavonic morphological tagset.

The resulting tagset 12 main categories: noun, verb, adjective, pronoun, adverb, adposition, conjunction, numeral, particle, interjection, abbreviation, and residual, having 0-10 attributes each. In total giving 156 attribute-value pairs

5. Evaluation

- The disambiguated portion of the RNC (about 5 million words)
- Three statistical POS taggers: TnT, TreeTagger and SVM Tagger
- 10% of the corpus held out for testing the performance of the taggers
- an experiment with ten second year British students of Russian revealed that intermediate level students are not able to spot the errors produced by the taggers: they too seem to analyze forms in isolation.
- Common problems: accusative vs. nominative; genitive singular vs. nominative/accusative plural for nouns; genitive singular vs. instrumental for feminine adjectives

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%

Overall accuracy of TnT, TT, and SVMTool, full tagset

	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

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

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

	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

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

Resources: `http://corpus.leeds.ac.uk/mocky/`

LREC 2008, 6th Language Resources and Evaluation Conference, 28-29-30 May 2008, Marrakech (Morocco).