SynTagRus (Russian National Corpus)

- Over 52,000 sentences as of 2012
  - from texts of a variety of genres(contemporary fiction, popular science, newspaper etc. from 1960-2012)
- A Sub-Corpus of the NRC
- Developed and maintained by the Laboratory of Computational Linguistics (LCL) in Moscow
- main purpose of the corpus is to facilitate academic research on the lexicon and grammar of a language, as well as the subtle but constant processes of language change within a relatively short period of time: from one to two centuries.
- Annotation consists of:
  - Morphological marking, syntactic tagging

Syntactic and Morphological annotation

- Done semi-automatically
  - First processed by ETAP-3 parser
  - Then manually corrected by linguists

- As Russian is a free word order language:

1) Relies on the Meaning-Text theory by Igor Melcuk

2) Uses a dependency tree

  - **Nodes:** the lemma, POS, morphological features (Eg. Aspect, tense, person, gender etc.);
  - **Arcs:** syntactic relations

- Uses a morphological dictionary with over 130,000 entries

Lexical Semantic and Lexical Functional Annotation

- SynTagRus currently only contains partial lexical functional annotation (Eg. Collocations)

- SynTagRus currently only shows lemmas of words occurring in the texts
  - does not disambiguate… ambiguous words, unless they have different lemmas or different POS tags.

- There is an ongoing project to add these into the corpus

Usage

- As a benchmark in regression tests designed to ensure stable performance of the ETAP-3 Russian Parser(Iomdin et al., 2012)
- As well as to refine the ETAP-3 Parser (Boguslavsky et al., 2011)
- And train other parsers(Shelmanov & Smirnov, 2014)
- As a source for the creation of statistical parsers for Russian(Nivre et al., 2008)

References

Boguslavsky, I., Iomdin, L., Timoshenko, S. P., & Frolova, T. I. (2009, April). Development of the Russian Tagged Corpus with Lexical and Functional Annotation. In *Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia* (pp. 83-90).Development of a dependency Treebank for Russian and its possible applications in NLP

Boguslavsky, I., Iomdin, L., Sizov, V., Tsinman, L., & Petrochenkov, V. (2011). Rule-based dependency parser refined by empirical and corpus statistics. In *Proceedings of the International Conference on Dependency Linguistics* (pp. 318-327).

Iomdin L. (2012). *Automatic text processing and deeply annotated text corpora of Russian: interaction and mutual impact* [PowerPoint slides]. Retrieved from http://korpus.sk/files/roadshow2012/Iomdin-Syntagrus.pdf

Iomdin L., Petrochenkov V., Sizov V., Tsinman L. (2012). ETAP parser: state of the art. Retrieved from http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Iomdin.pdf

Nivre, J., Boguslavsky, I. M., & Iomdin, L. L. (2008, August). Parsing the SynTagRus treebank of Russian. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 641-648). Association for Computational Linguistics.

Shelmanov A. O., Smirnov I. V. (2014). Methods for Semantic Role Labeling of Russian Texts. Retrieved from http://www.dialog-21.ru/digests/dialog2014/materials/pdf/ShelmanovAOSmirnovIV.pdf

ISLRN

Title	**SynTagRus**
Full Title	SynTagRus Corpus
Resource Type	Corpus
Source/URL	http://www.ruscorpora.ru/en/search-syntax.html
Format/MIME Type	text/xml
Size/Duration	Over 52,000 Sentences
Access Medium	Online
Description	A sub-corpus of the Russian National Corpus, SynTagRus is a corpus of Russian Texts annotated with dependency-type syntactic structures, with full morphological and syntactic markup
Version	2?
Media Type	Text
Language	Russian, English
Resource Creator	Laboratory of Computational Linguistics of the Institute of Information Transmission Problems in Moscow
Distributor	Laboratory of Computational Linguistics of the Institute of Information Transmission Problems in Moscow
Rights Holder	Laboratory of Computational Linguistics of the Institute of Information Transmission Problems in Moscow