Automatic Conversion of Russian Treebank SynTagRus into PDT Style

David Mareček, Natalia Kljueva
Charles University in Prague, Institute of Formal and Applied Linguistics

Motivation:
- to incorporate Russian language into TectoMT framework
- to compare Russian tectogrammatical structures with already developed Czech and English structures
- usage in the Russian-Czech and Russian-English machine translation

Format Conversion

Tectogrammatical Analysis

Assigning Functors

Functor = semantic relation between a node and its parent
- simple rules which use syntactic relation, tag, and lemma of the node

Examples:
- appos → APPS; сравнит → CPR; примыкает → PAR; 1-компл → PAT
- preposition 'в' + case 'vinitelnyj' → DIR3
- 'чтобы' (to) or 'с целью' (with the aim of) → AIM

Different handling coordinations

SynTagRus:
- Meaning Text Theory
- First coordination member is the head and each other member (including conjunctions) is attached to the previous one.

PDT:
- Functional Generative Description
- Conjunction is the head and all coordination members are attached to it.

- necessary to rehang coordinations
- problems with common dependents

Function words
- hidden on the tectogrammatical layer
- attached to one of the content words
- edges to be collapsed are highlighted

Examples:
- preposition attached to the appropriate noun
- auxiliary verb 'быть' (in passives or in future tense) attached to the content verb
- subordinating conjunctions 'что' (that), 'чтобы' (so that), or 'потому что' (because) attached to the head of subordinated clause
- modal verb e.g. 'хотеть' (want), 'мочь' (can), 'надо' (should), or 'должен' (must) attached to the content verb
- ...

Assigning Functors

Function = semantic relation between a node and its parent
- simple rules which use syntactic relation, tag, and lemma of the node

Examples:
- appos → APPS; сравнит → CPR; примыкает → PAR; 1-компл → PAT
- preposition 'в' + case 'vinitelnyj' → DIR3
- 'чтобы' (to) or 'с целью' (with the aim of) → AIM

Tectogrammatical Analysis

Syntax relation

Morphological features

S = substantivum
ЕД = singular
ЖЕН = feminine
ВИН = case 'vinitelnyj'

SynTagRus statistics:

32,242 sentences
461,297 tokens (excluding punctuation)

texts from:
- journal articles (77.8 %)
- online news (12.4 %)
- fiction (9.8 %)