Converting Russian Treebank SynTagRus into Praguian PDT Style

David Mareček and Natalia Kljueva

marecek@ufal.mff.cuni.cz, kljueva@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics,
Charles University in Prague

Workshop on “Multilingual resources, technologies and evaluation for central and Eastern European languages”

Borovets, Bulgaria, September 17, 2009

Motivation

- Besides the Czech treebank (Prague Dependency Treebank) we develop also treebanks of other languages using the same annotation scheme
  - PDT – Prague Dependency Treebank (Czech)
  - PEDT – Prague English Dependency Treebank (Wall Street journal)
  - PADT – Prague Arabic Dependency Treebank
  - PCEDT – Prague Czech-English Dependency Treebank (parallel corp.)
- We want to add Russian
- Languages are easier to compare when annotated using the same annotation scheme

Outline

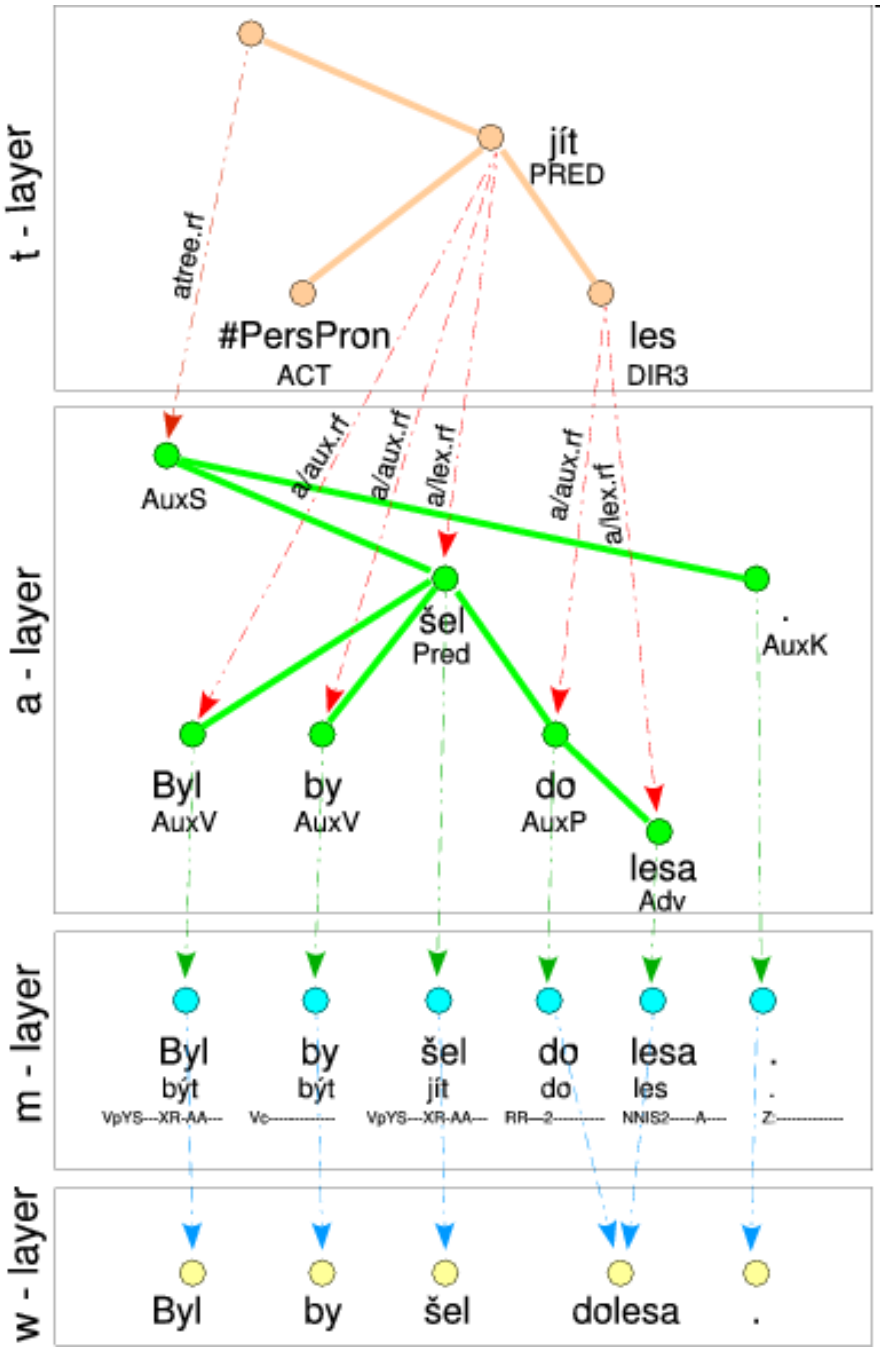
- Prague Depenency Treebank
- SynTagRus
  - Dependency treebank of Russian
- TectoMT software framework
  - Used for the conversion
- Conversion of SynTagRus into PDT style
  - Format conversion
  - Handling coordinations
  - Function words
  - Functor assignment
- Small Czech-Russian parallel treebank
- Conclusions and future work

Prague Dependency Treebank

- Dependency treebank of Czech
- Consists of three interlinked annotation layers
  - Morphological
  - Analytical (syntax)
  - Tectogrammatical (deep syntax)
- 115,000 sentences and 2,000,000 tokens (including punctuation) from newspapers and scientific journals
  - All of them are annotated on the morphological layer
  - 75% also on the analytical layer
  - 45% on all three layers

PDT Layers

- Morphological layer
  - Lemma and tag are assigned to each token
- Analytical layer
  - Dependency tree, where each node corresponds to one token
  - Syntactic functions assigned
- Tectogrammatical layer
  - Dependency tree, where only content words have their own nodes
  - Words dropped on the surface are added
  - Many attributes assigned to nodes (functor, semantic POS, semantic morphological categories…)

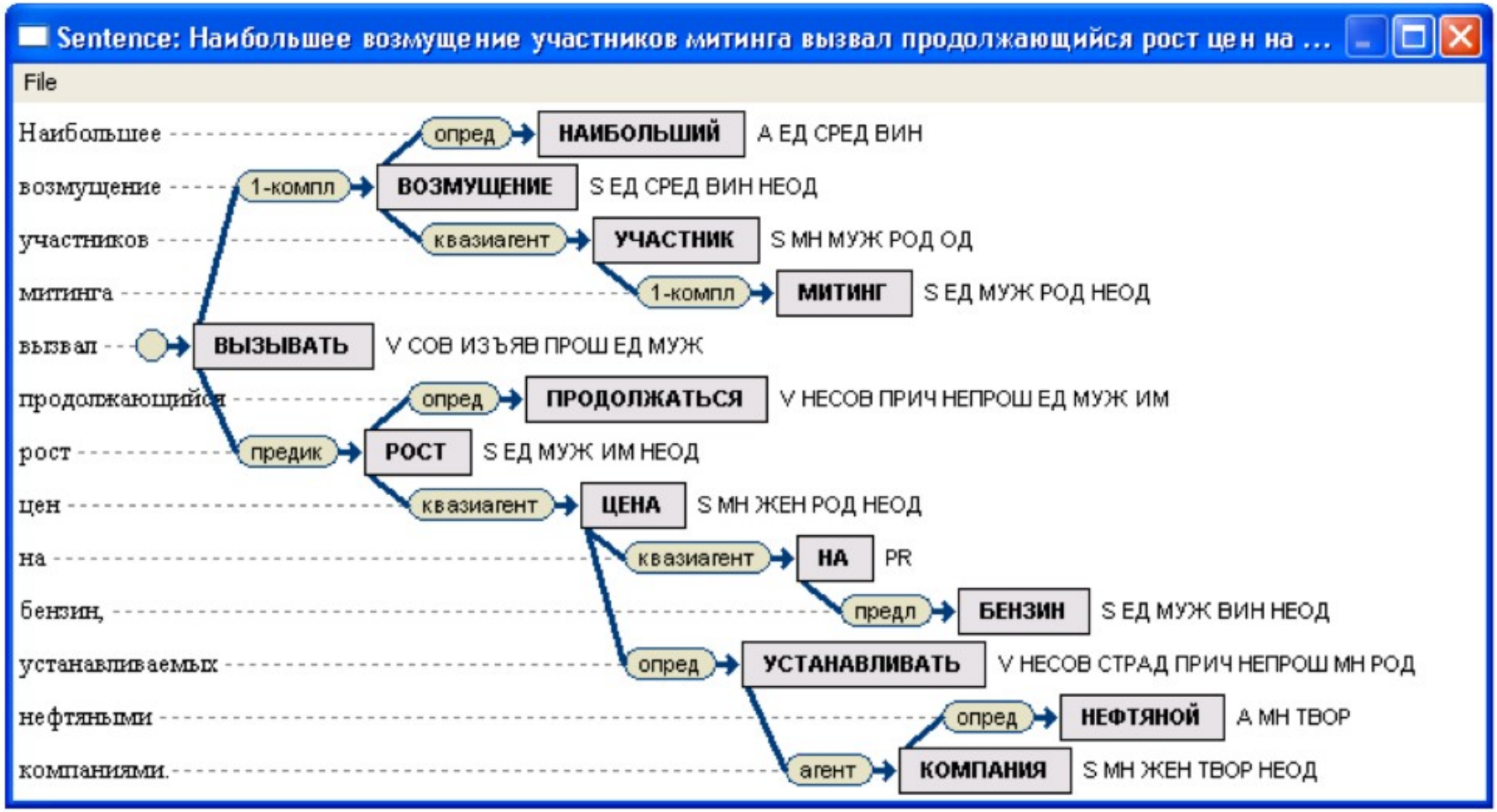


SynTagRus

- Syntactically annotated corpus of Russian
  - Developed in Institute for Information Transmition Problems on Russian Academy of Sciences
- Text sources:
  - Uppsala University Corpus of Contemporary Russian prose
  - Newspaper articles
- Statistics:
  - 32,000 sentences
  - 460,000 words

Example of sentence annotated in SynTagRus

- .



SynTagRus - Attributes

- Word Form
- Lemma
- Tag
  - Part of Speech (S, A, V, PR, …)
  - Set of morphological features (number, gender, case, person, tense, …)
- Dependency type (syntactic relation between the node and its parent)
  - ‘предик’ - between a verb and the subject
  - ‘1-компл’ - between a verb and its direct object
  - ‘предл’ - between a preposition and a noun
  - … and many others

TectoMT software framework

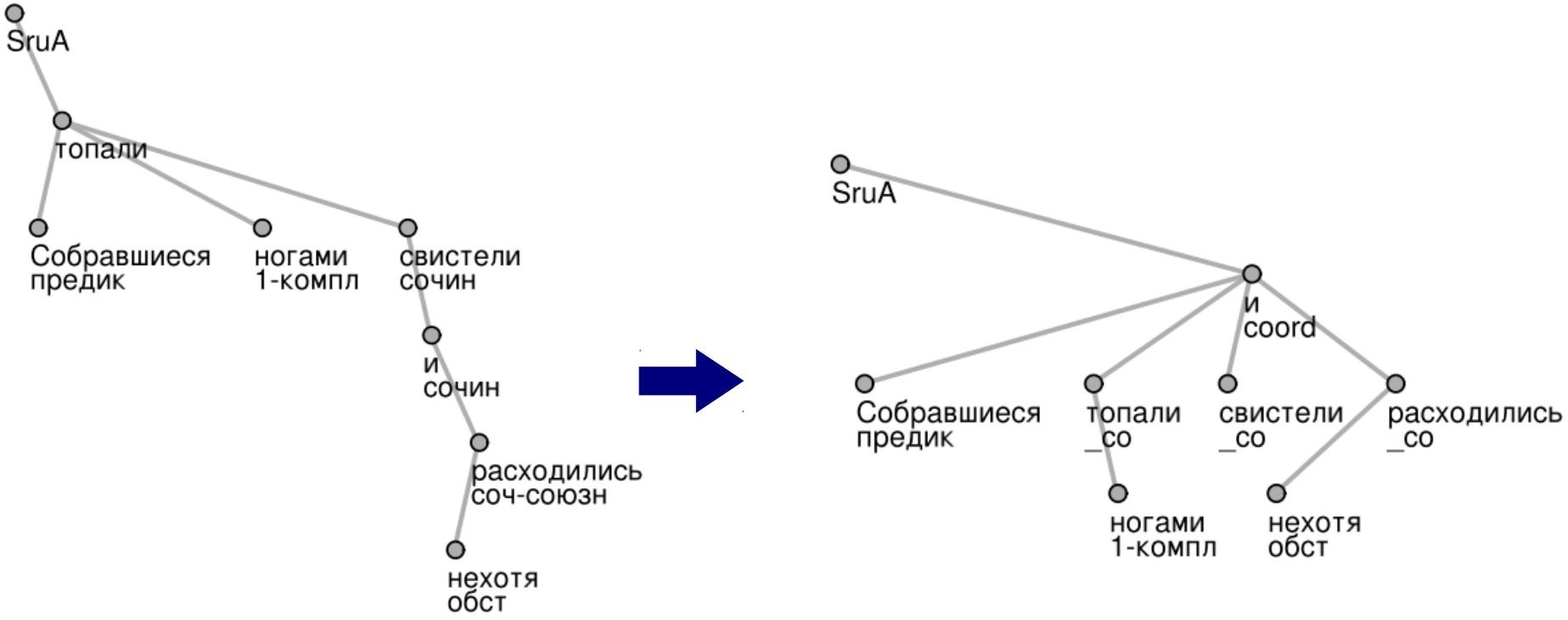
- Consists of many linguistics tools
  - Tokenizers, taggers, dependency parsers, constituent parsers
  - Named entity recognizers, tools for bilingual alignment
  - Various analysis and synthesis tools
- Perl interface
  - Tools programmed in other languages are wrapped into a Perl script
- One common data format ‘TMT’, which is based on XML
  - Format convertors to/from TMT (e. g. tmt_to_txt, syntagrus_to_tmt)
- Tree viewer and editor ‘TrEd’
  - We can view and edit different types of trees (constituent, dependency), their attributes, links between them, …
- Originally the framework was developed mainly for Czech and English, now we add other languages (Arabic, German, Russian) and the framework becomes more language-independent.

Format Conversion

- SynTagRus annotation covers all the features that are necessary to build morphological and analytical layer
  - It was only technical problem to convert SynTagRus XML schema into our TMT schema and build the firs two layers for every Russian sentence
  - After the conversion, the TectoMT framework can be used.

Handling coordinations

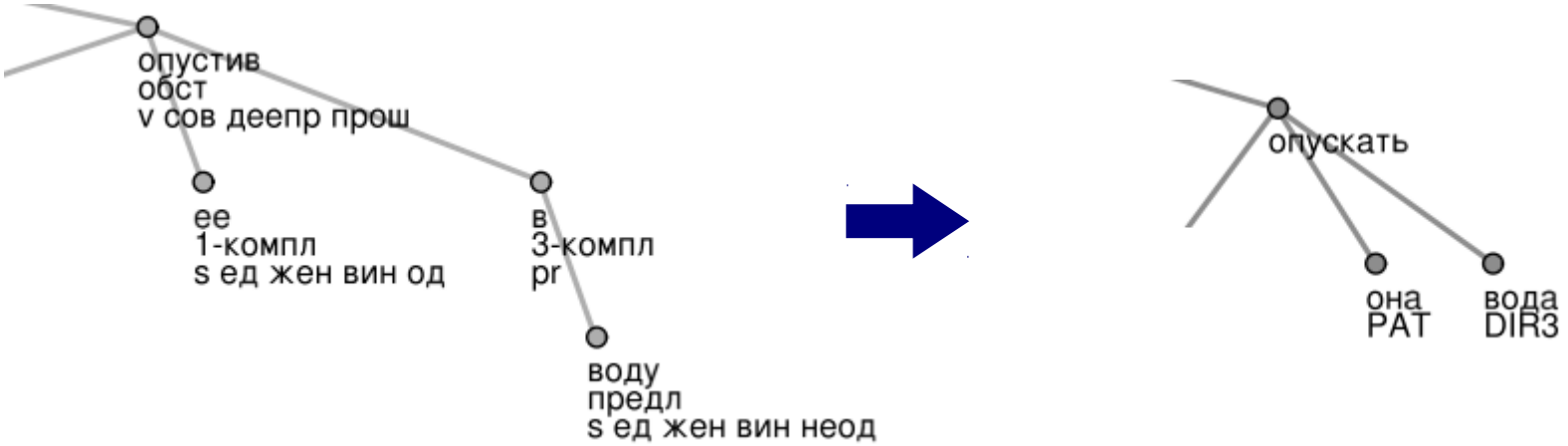
- SynTagRus
  - Meaning Text Theory
- PDT style
  - Functional Generative Description



Собравшиеся топали ногами, свистели и нехотя расходились.
[People stamped their feet, whistled and left unwillingly.]

Building the tectogrammatical layer

- Now we have already the Russian analytical trees and we want to build the tectogrammatical trees – the deep syntactic representation
  - Function words (prepositions, auxiliary verbs, modal verbs, …) will not have their own node in the tectogrammatical tree
  - The meaning of the function words will be expressed by functors and grammatemes (the attributes of respective content-word nodes).
- Three steps:
  - Assign function words to the content ones in the analytical tree
  - Build the tectogrammatical tree
  - Assign attributes to its nodes



Rules for functional word assignment

- ***prepositions*** – A preposition is assigned to its child node (a noun), if the syntactic relation is ‘предл’ (prepositional).

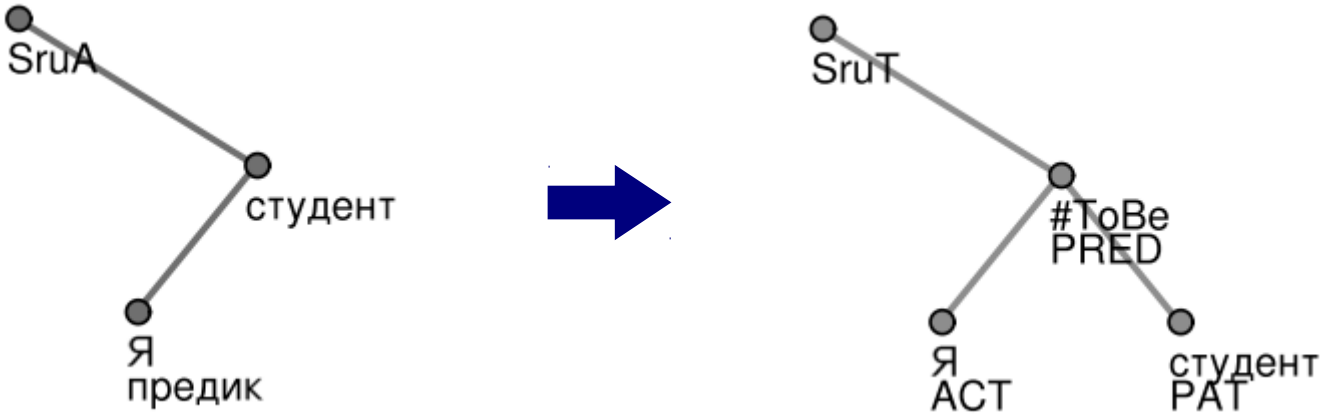
- ***subordinated conjunctions*** – Conjunctions ‘что’ (that), ‘чтобы’ (so that), or ‘потому что‘ (because) are assigned to their child nodes, if the syntactic relation between them is ’подч-союзн’ (subordinate clause with conjunction).

- ***modal verbs*** – A verb which lemma is ‘хотеть’ (want), ‘мочь’ (can), ‘надо’ (should), or ‘должен’ (must) is assigned to its child node, if the child node is verb in infinitive form.

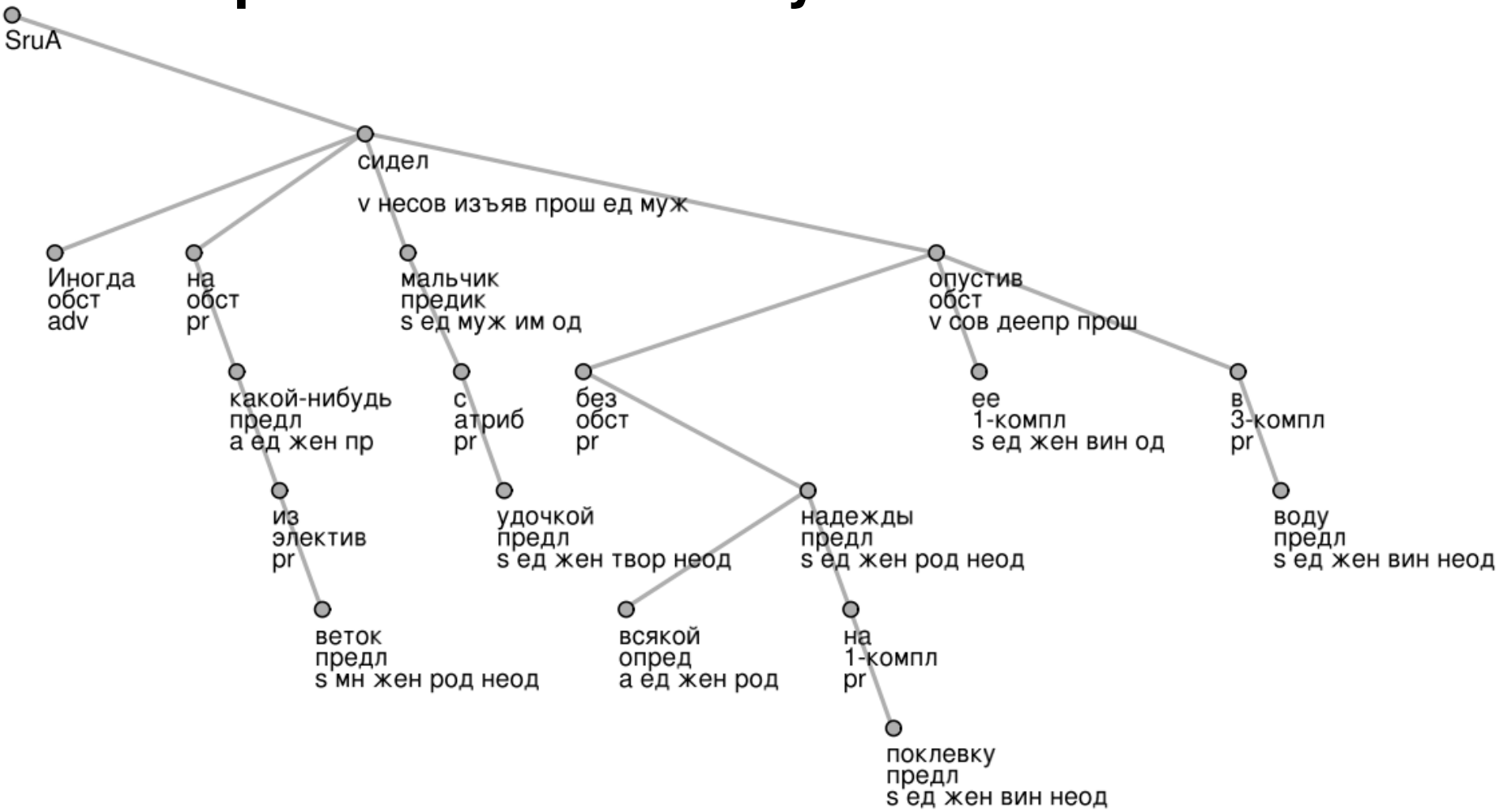
- … and others

Elided ‘to be’ in Russian

- Some words (dropped on the surface) are added into the tectogrammatical trees
  - In this example it is the verb ‘to be’ in Russian



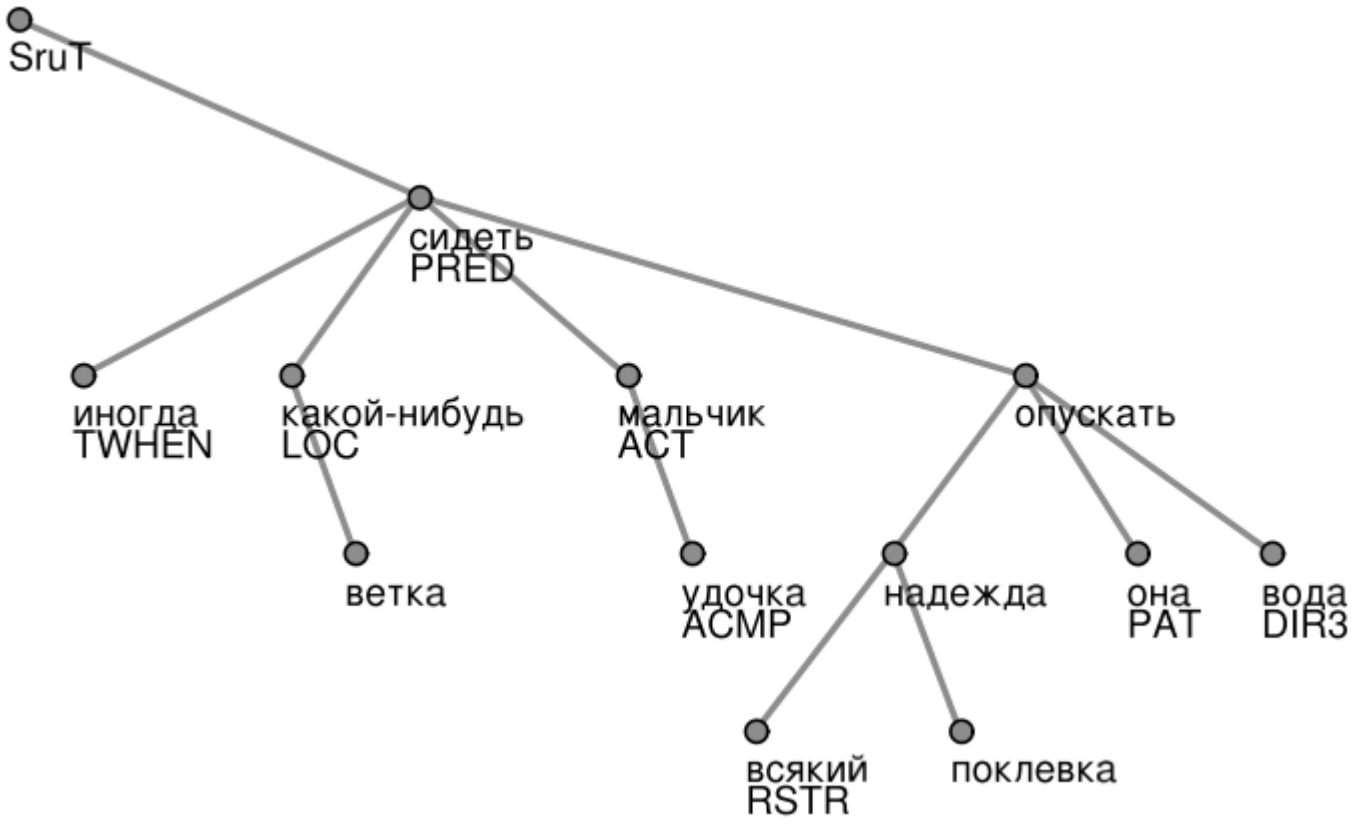
Example of Russian analytical tree



Иногда на какой-нибудь из веток сидел мальчик с удочкой, без всякой надежды на поклевку опустив ее в воду.

[Now and than a boy with a fishing rode was sitting on a branch, dropping it into the water without any hope to catch fish.]

Example of Russian tectogrammatical tree



Иногда на какой-нибудь из веток сидел мальчик с удочкой, без всякой надежды на поклевку опустив ее в воду.

[Now and than a boy with a fishing rode was sitting on a branch, dropping it into the water without any hope to catch fish.]

Conclusion and Future Work

- We described the first steps of converting the Russian dependency treebank SynTagRus into the PDT style and developing tectogrammatical layer for Russian
  - We are on half of the way
- In the future, we plan to continue with adding more (often more complex) rules for assigning functors and other attributes.
- Experimenting with the Czech-Russian parallel treebank
  - Some chunks was translated into Czech and automatically parsed up to the tectogrammatical trees.
  - Czech and Russian tectogrammatical representations of a sentence are much more similar than their surface shapes.

Thank you for your attention