Towards Parallel Czech-Russian Dependency Treebank

Natalia Klyueva and David Mareček

Charles University in Prague
Institute of Formal and Applied Linguistics,
E-mail: {kljueva, marecek}@ufal.mff.cuni.cz

Abstract

In this paper we describe initial steps in constructing a Czech-Russian dependency treebank and discuss the perspectives of its development. Following the experience of the Czech-English Parallel Treebank we have taken a syntactically annotated “gold standard” text for one language (Russian) and run an automatic annotation on the respective parallel text for the other language (Czech). Our treebank includes also automatic word-alignment.

1 Introduction

Large number of treebanks has appeared recently, and constructing the parallel treebanks is becoming more popular. This type of linguistic data presents valuable resource for both theoretical research in comparative syntax and NLP applications like Machine Translation. Parallel treebanks are generally compiled for English and some other language, but exceptions exist. To the best of our knowledge, no such parallel treebank exists for related Slavic languages.

We have created a small parallel treebank using data and tools from two existing treebanks. The manually annotated Russian data are taken from SynTagRus treebank [8]. Tools for the parsing the corresponding text in Czech are taken from the TectoMT framework [10]. We believe that our parallel treebank will open a road to the development of such treebanks for other Slavic languages.

Our project is connected to PCEDT - the Prague Czech-English Dependency Treebank [2]. Data for an annotation in it were taken from the Penn Treebank, precisely, a part which contains the texts from the Wall Street Journal. Though our project can not be compared to PCEDT in both quality and quantity, as the translation from English into Czech was made as closely to the original as possible, and the size of it is suitable for NLP tasks, for example Machine Translation.

As in PCEDT, we borrowed the text annotated within another framework and transformed it into a PDT style. It was easier for us because both treebanks annotate dependency structure, not phrase structure. Still, we were not able to manually check the automatically parsed Czech text as it is done in the PCEDT.

Another very similar project is SMULTRON [1], the English-German-Swedish multilingual treebank, which also disposes a set of tools, as for example the Tree Aligner, that may be useful for our Treebank in the future.

This paper is structured as follows. Section 2 provides an overview on the two treebanks - the SynTagRus for Russian and the PDT for Czech, here we also introduce the data we chose for our treebank. In Section 3 an adaptation of the Russian annotation schema to the PDT style is described. Section 4 demonstrates the process of an automatic annotation of Czech text. Section 5 overviews the core – compilation and description of the treebank. Section 6 provides an example of a treebank exploitation. Finally, we conclude in Section 7.

2 Data and Tools

2.1 PDT and TectoMT

We decided to choose the Prague Dependency Treebank as a platform for our treebank, as it is more experienced with a parallel treebank handling and dispose tools for this. PDT contains 115,844 sentences from newspapers and journals.

In Prague Treebanking school a sentence is annotated on three layers: morphological, analytical and, tectogrammatical.

2.1.1 The Morphological Layer

Each word in a tree is represented as a node with a lemma and a tag assigned. The morphological tag is so-called positional, 15 positions are filled with an appropriate morphological category (Part of Speech, Gender, Number, Case, Person, Tense, etc.). All the sentences in PDT are annotated on this level.

2.1.2 The Analytical Layer

Syntactic annotation is presented in form of dependency tree, where each morphologically annotated token from the previous level becomes a node with an assigned analytical function. Analytical function (afun) reflects a syntactic relation between a parent and a child node and is stored as an attribute of the child. Examples of an analytical functions: Subject (Sub), Predicate (Pred), Object (Obj) etc. Analytical layer is annotated in 75 % of PDT texts.

2.1.3 The Tectogrammatical Layer

The annotation on the tectogrammatical layer (t-layer) goes deeper towards the level of meaning. Function words (prepositions, auxiliary verbs, conjunctions, etc.) are removed from the correspondent analytical tree and are stored as attributes of autosemantic words, leaving only content words as the nodes on the t-layer. Tectogrammatical layer is annotated on 45 % of PDT texts.

The tools for automatic annotation of Czech sentences on these three layers are freely available in a TectoMT framework [10], which is used in our work too.

2.2 SynTagRus

SynTagRus is a collection of texts annotated on a morphological and a deep syntactic level. Texts in SynTagRus are mainly newspaper articles with a small amount of modern prose texts, it contains approximately 460,000 words. The treebank is coded in an XML-based schema.

Words are represented by nodes, which have three morphological attributes: word form, lemma and tag. Unlike a Czech positional tag, where a morphological feature has its own fixed position, the tags for Russian are conditional - the sequence of features depends on the part of speech. This difference, however, is not relevant to us as we leave the morphological tags untranslated, focusing rather on the syntax and the deep syntax transfer.

The nodes are connected between each other with the arcs that are marked with one of 78 syntactic relations (Predicative, Attributive, Adverbial etc.) One of the main “surface” differences from PDT is that the SynTagRus does not regard punctuation marks as nodes, whereas in the PDT analytical (syntactic) level punctuation symbols have even their own syntactic function.

2.3 Data for our experiment

For a parallel treebank we have chosen a part of a Russian novel “Kafedra” (“The Faculty”) by I. Grekova, because this novel was also translated into Czech and 480 sentences of it were annotated within the SynTagRus. Those sentences formed the core of our treebank. Probably more sufficient from a point of view of sentence correspondence will be translations of news articles, but they do not exist. We disposed only the printed version of the book which we scanned and aligned the sentences in the text manually.

The main challenge to handle the corpus is its novel translation into Czech. A sentence translated into Czech sometimes bears only a meaning of a source Russian sentence, and it is rather difficult to make the word alignment.

This problem is also multiplied by free word-order of those two Slavic languages. First we supposed that this common syntactic feature will contribute to the similarity of sentences. Afterwards we have found out that while translating the free word order Czech construction, in the Russian sentence the words can be mixed up in another way.

3 Format Transfer for Russian data

SynTagRus is coded in an XML-based format, which we transformed into a PML (Prague Markup Language) format. It would be also rather straightforward to transfer Russian morphological tags into Czech. Morphological systems of the two languages are almost similar, both Czech and Russian have the same cases except for Vocative in Czech, verb tense system is also very close.

On the other hand if we want to be consistent, we should also make a transformation of syntactic properties (Russian syntactic functions) into afuns (Czech analytical functions). Here we face a big problem, because the two annotation schemes have different principles of annotation in this case. There are more than 78 syntactic functions in SynTagRus and only 28 afuns on the analytical layer in PDT, most of which can be mapped into those from SynTagRus (Predicative, Adverbial, Auxiliary relations).

Still, we should not forget about th information on the tectogrammatical layer for Czech, or functors. We argue, that the combination of an analytical function and a functor for Czech can correspond in some cases to a syntactic function from SynTagRus. In other words, the syntactic layer of annotation for Russian is more deep and semanticalized, and it is one layer. Whereas the Czech annotation draws a distinction line between syntax and semantics, leaving syntactic features to the analytical layer and semantics to the tectogrammatical one. This fact and some possible solutions of this problem can be illustrated by an example of a verb argument structure. For instance, in SynTagRus the complement relations are described as syntactic functions “n-compl”, where n is a sequence number of an actant. In the Czech PDT it can correspond to either tectogrammatical functor “Patient” or “Means”. In order to capture such differences we wrote a set of rules, for example they can be schematized as follows:

```
Ru:  1-compl in Accusative case → Cz:  Patient,
Ru:  1-compl in Instrumental case → Cz:  Means.
```

The rules of transfer are now currently under development, and we have found corresponding functors in Czech for all syntactic relations in Russian. More information on the format transfer between the treebanks can be found in [4].

4 Parsing the Czech text

One of the biggest challenges of our work was to annotate the raw Czech data on all the levels - morphological, syntactic and a bit semantic, so that these sentences can be "comparably" aligned to their high-quality manually annotated Russian counterparts. Translated Czech sentences were automatically analyzed using TectoMT framework [10].

The following steps were done:

- tokenization,
- tagging and lemmatization using Morce tagger [12],
- parsing with McDonald's MST parser [5],
- automatic conversion to tectogrammatical trees using mainly rule-based scripts, which are included in TectoMT framework.

Obviously, mistakes in automatically parsed Czech trees occurs. The unlabeled accuracy of the Czech parser is about 85%. We plan to fix them manually in the future.

5 Parallel Treebank Compilation

The parallel treebank is represented on three layers: morphological, analytical and tectogrammatical. The size of the treebank is not very big in comparison with treebanks mentioned in Section 1, and we are currently looking for ways to enlarge the corpus. The statistics of our parallel treebank is summarized in the Table 1.

Table 1: Summary of the Treebank size

	Czech	Russian
sentences	480	480
words	5131	5895

Trees are visualized in TrEd editor[1], which is used for the annotation of the PDT. A screenshot of the annotation on all three layers for both Czech and Russian sentences can be seen in Figure 1. Now we will briefly describe annotation layers of the parallel treebank.

5.1 The Morphological layer and the Word Alignment

The morphological layer shows a sentence in Czech and Russian, where the words go in a linear manner, and they have their morphological properties attached. The whole corpus is automatically aligned on the level of words. For this purposes we ran the GIZA++ tool [9] on parallel texts lemmatized both on the Czech and Russian side. The two resulting one-directional alignments were then symmetrized using intersection symmetrization. For better alignment results we added to our small parallel data the Czech-Russian part of parallel corpus UMC [3]. On the sample of 100 sentences we made a manual evaluation of a word alignment quality,

[1] http://ufal.mff.cuni.cz/ pajas/tred/

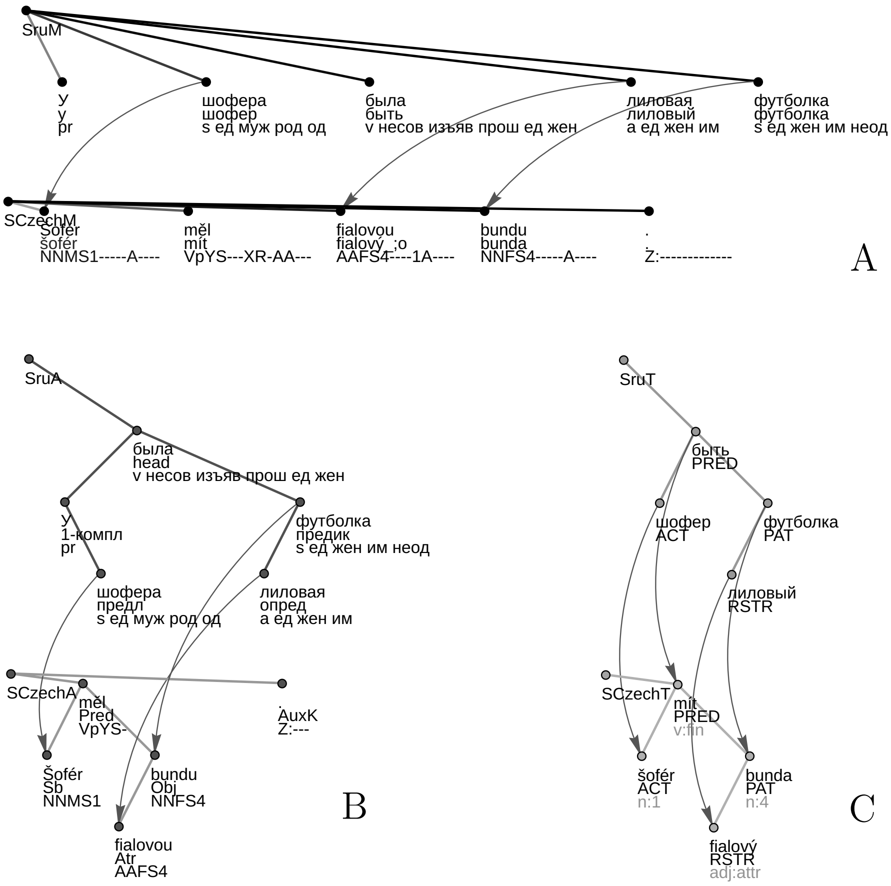

Figure 1: Representation of the sentence "The driver had a lilac coat" for Russian (at the top) and Czech (at the bottom) on morphological layer (A), analytical layer (B) and tectogrammatical layer (C).

its precision reached 85 %. In the future, we plan to improve the word alignment by introducing a good Czech-Russian dictionary.

5.2 The Analytical Layer

The core goal of this project is a task of annotation of the treebank at least on the analytical level, so that syntactic correspondences between the languages can be seen. If not taking into account some surface incorrespondences in Czech and Russian trees caused by different annotation scheme, as, for instance, punctuation marks in Czech scheme which are ignored in SynTagRus, we can compare syntactic constructions in both languages. Figure 2 illustrates a sentence which has more or less similar syntactic structure, and the shapes of two trees are evidently close. In the next section we will show an example of trees with a different syntactic structure.

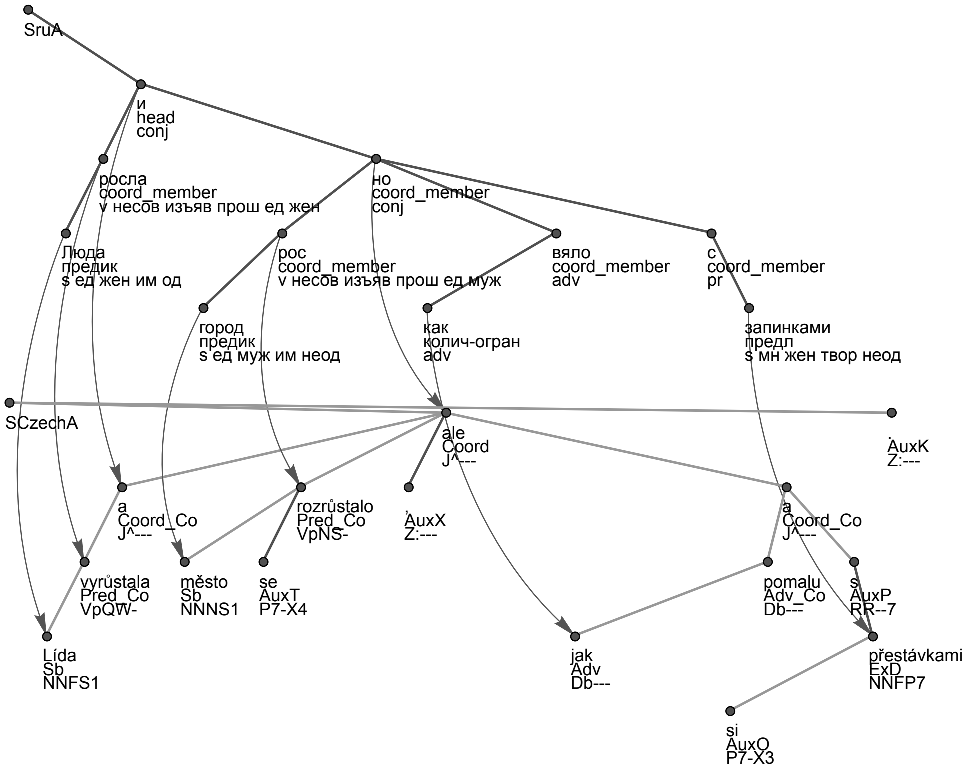

Figure 2: Aligned analytical representation of the sentence (lit.)"Lida was growing up and the town was growing, but somehow slowly, with breaks".

5.3 The Tectogrammatical Layer

The tectogrammatical layer of our parallel treebank is so far annotated only preliminary. It would be a huge work to make correspondences between Czech functors and Russian analytical functions. Still, tectogrammatical trees in two languages will be more similar, than the corresponding analytical trees. One of our tasks for future will be improving the tectogrammatical annotation for this treebank. First insight into the tectogrammatical annotation of Russian is described in [4].

6 Sample Analysis of a Sentence

We have described the preliminary research of how the Czech-Russian treebank can look like. Due to the small size of the parallel treebank it can not be used for the purposes of Statistical Machine Translation, as the PCEDT. However, this annotated data on each of the three layers can bring some insight into the comparative studies that can be useful while designing a Rule-Based or Hybrid Machine Translation system between the languages. As an example of such exploiting for differences that serve as a basis for MT rules of transfer, we will examine the sentence from the Figure 1.

The *morphological annotation* will provide evidence on whether or not sentences in two languages consist of words with the same or different part of speech, and how similar the morphological properties of those words are. In our example there are four lemmas in Czech and five in Russian (extra one is a preposition).

The *syntactic annotation* can help while inducing basic rules of the syntactic transfer for the Rule-Based MT system. For example, a frequent possessive construction with the verb "to have" in Czech and "to be" in Russian depicted as a tree reflects a difference, which is a candidate for a syntactic rule. To continue, in Czech and Russian sentences the same aligned words have different syntactic functions (“driver” - Subject and a child of the “verb” in Czech, Object and a “child” of the preposition in Russian).

Lastly, two trees on the *tectogrammatical layer* are identical and the corresponding nodes have the same tectogrammatical functors, as this level of annotation stands closer to the “Interlingua”.

7 Conclusion and Future Work

We have shown here the initial phase of building the Czech-Russian Dependency Treebank. On the small sample of the data we made the preliminary correspondence between the two annotation schemes, which will be useful while adding new data to the treebank. One of the possible directions of our research is also making use of automatic annotation tools from the SynTagRus - the tagger and the parser - so that we can annotate a parallel corpus of Czech and Russian languages on syntactic level, not being dependent on the data from the monolingual treebanks. This will enlarge our corpus size at the price of quality, because in addition to the Czech parser mistakes, there will be also mistakes from the Russian parser. The treebank described is not published on-line because of the copyrights. Still, it will be widely exploited for the internal research purposes, namely for constructing rules for the RBMT system between Czech and Russian.

8 Acknowledgments

The research was funded by grants MSM 0021620838, GA201/09/H057, and GAUK 116310/2010. We are also grateful to Leonid Iomdin who provided the Russian data for the project and to the anonymous reviewers for their valuable remarks.

References

[1] Gustafson-Capkova, Sofia, Samuelsson, Yvonne, and Volk, Martin (2007) SMULTRON (version 1.0) - The Stockholm MULtilingual parallel TRee-

bank. http://www.ling.su.se/dali/research/smultron/index.htm. An English-German-Swedish parallel Treebank with sub-sentential alignments.

[2] Čmejrek, Martin, Cuřín, Jan, Havelka, Jiří,Hajič, Jan and Kuboň, Vladislav (2004) Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation *In 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.*

[3] Klyueva, Natalia and Bojar, Ondřej (2008) UMC 0.1: Czech-Russian-English Multilingual Corpus, *Proceedings of International Conference Corpus Linguistics*, Saint-Petersburg, pp. 188–195.

[4] Mareček, David and Kljueva, Natalia (2009) Converting Russian Treebank SynTagRus into Praguian PDT Style, *Proceedings of Multilingual resources, technologies and evaluation for Central and Eastern European languages*, Borovets, Bulgaria.

[5] McDonald, Ryan, Pereira, Fernando, Ribarov, Kiril and Hajič, Jan (2005) Non-Projective Dependency Parsing using Spanning Tree Algorithms, *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP)*, pp. 523–530, Vancouver, BC, Canada.

[6] Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993) Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics 19*, pp. 313–330, [Reprinted in Armstrong, Susan (ed.) (1994) *Using large corpora*, pp. 273–290. Cambridge, MA: MIT Press.]

[7] Mel'čuk, Igor (1988) Dependency Syntax: Theory and Practice, State University of New York Press.

[8] Nivre, Joakim, Boguslavsky, Igor and Iomdin, Leonid (2008) Parsing the SynTagRus Treebank, *Proceedings of COLING08*, pp. 641–648.

[9] Och, Franz Josef and Ney, Hermann (2003) A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics,* **1**, 29, pp. 19–51.

[10] Popel, Martin, Žabokrtský, Zdeněk (2010) TectoMT: Modular NLP Framework, *Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, pp. 293–304.

[11] Sgall, Petr, Hajičová, Eva and Panevová, Jarmila (1986) The Meaning of the Sentence in Its Pragmatic Aspects, Reidel.

[12] Spoustová, Drahomíra, Hajič, Jan, Votrubec, Jan, Krbec, Pavel and Květoň, Pavel (2007) The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech, *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, Praha, pp. 67–74.