Converting Russian Treebank SynTagRus into Praguian PDT Style

David Mareček and Natalia Kljueva
Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00, Praha 1, Czech Republic
marecek@ufal.mff.cuni.cz kljueva@ufal.mff.cuni.cz

Abstract

In this paper, we report a work in progress on transforming syntactic structures from the SynTagRus corpus into tectogrammatical trees in the Prague Dependency Treebank (PDT) style. SynTagRus (Russian) and PDT (Czech) are both dependency treebanks sharing lots of common features and facing similar linguistic challenges due to the close relatedness of the two languages. While in PDT the tectogrammatical representation exists, sentences in SynTagRus are annotated on syntactic level only.

Keywords

Dependency treebank, tectogrammatical trees, dependency relations, parallel corpora

1 Introduction

Treebanking in Prague comprises not only the annotations of Czech. Besides the main project of Prague Dependency Treebank (PDT) [3], there are several other projects using the same schema for annotating other languages. We should mention the Prague Arabic Dependency Treebank (PADT) [4] and Prague English Dependency Treebank (PEDT) [1], which contains texts from Wall Street journal manually annotated in the PDT style. The Prague Czech-English Dependency Treebank (PCEDT) [2] was developed by translating PEDT into Czech and annotating it also on the Czech side.

Our goal is to convert the Russian corpus SynTagRus [7] into the PDT annotation scheme and build the tectogrammatical (deep-syntactic) layer for Russian. We also develop a small Russian-Czech parallel treebank so that we can compare the two closely-related languages and study structural similarities and differences, which could be useful for developing machine translation systems.

2 Description of the treebanks

2.1 Prague Dependency Treebank

Prague Dependency Treebank (version 2.0) [3] is a treebank of Czech, which consists of three interlinked annotation layers: the morphological layer, the analytical layer (describing the surface syntax) and the tectogrammatical layer (describing the deep syntax – transition between syntax and semantics). A highly simplified example of the annotation layers is in Figure 1. The theoretical background of PDT has its roots in the Prague School of Functional and Structural Linguistics, and especially in the language description framework called Functional Generative Description [9]. The following paragraphs summarize the main features of the three layers.

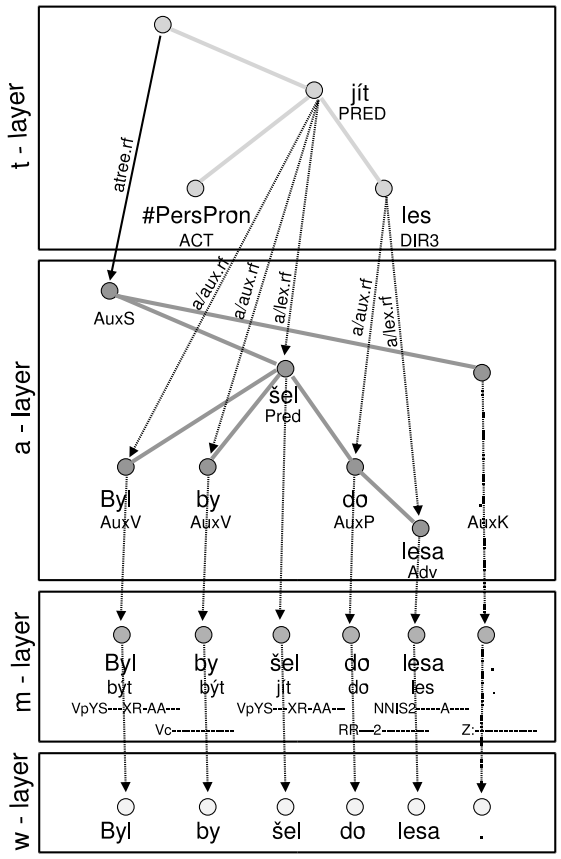

Figure 1: *PDT 2.0 annotation layers (and the layer interlinking) illustrated (in a simplified fashion) on the sentence "Byl by šel do lesa." ([He] would have gone into forest.)*

Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages 2009, Bulgaria, pages 26–31

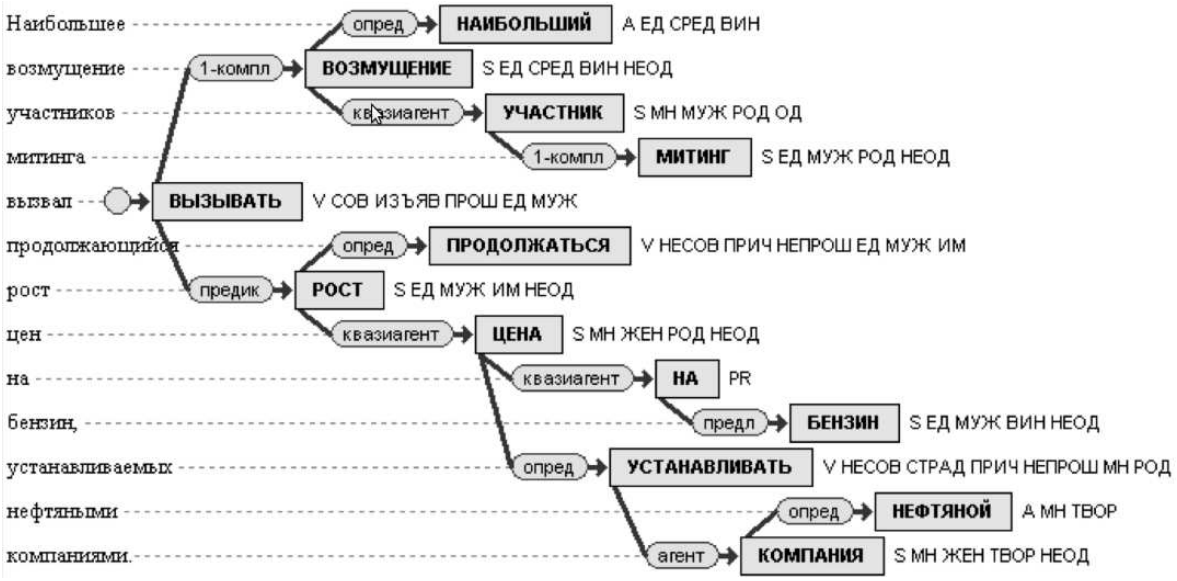

Figure 2: *A syntactically annotated sentence from the SynTagRus treebank. Lemmas in rectangles are followed by tags, syntactic relations are in ovals.*

At the morphological layer (m-layer), a sentence is divided into tokens (words, punctuation marks, and other symbols). Lemma and positional morphological tag are assigned to each token.

At the analytical layer (a-layer), a rooted dependency tree is being build for every sentence. Every token from the morphological layer becomes exactly one node in the analytical tree. Only one node – the "technical" root – is added. An analytical function (such as Subject, Object, Attribute) is assigned to each node, but in fact it captures the type of dependency relation between the given node and its parent node. However, there are also edges representing non-dependency relations (e.g. in coordination structures).

At the tectogrammatical layer (t-layer), each sentence is represented as a complex deep-syntactic dependency tree (tectogrammatical tree), in which only autosemantic words have nodes of their own. Functional words like prepositions, subordinating conjunctions, auxiliary verbs, and modal verbs are represented in the respective nodes in the form of their attributes. On the other hand, tectogrammatical trees contain nodes that have no counterparts in the surface shape of the sentences, for instance nodes corresponding to 'pro-dropped' subjects. Each node has its tectogrammatical lemma, functor (which determines the type of semantic relation between the node and its parent), semantic part of speech, grammatemes (semantically-oriented counterparts of morphological categories such as aspect, degree of comparison, modality, gender, iterativeness, negation, number, person, or tense).

The corpus contains 115,844 sentences (1,957,247 tokens including punctuation and other special characters) from newspapers and scientific articles. All of them are annotated on the m-layer, 75% on the a-layer and 45% on all three layers.

2.2 SynTagRus

SynTagRus is a syntactically annotated corpus of Russian based on the theory "Meaning-Text" [6]. In SynTagRus, sentences are represented as trees, in which words are nodes and edges between them are marked with the appropriate syntactic relation. Unlike in PDT, punctuation marks are not annotated in SynTagRus. They are included, but do not carry any labeling and they are not included in syntactic trees. An annotated sentence from SynTagRus is depicted in Figure 2.

Each word (node in a tree) has five attributes in the SynTagRus XML format:

- *id* – linear position of the word in the sentence,
- *dom* – id of its parent node,
- *lemma* – morphological lemma,
- *feat* – morphological tag.[1] Part of speech at the first position is followed by a sequence of respective features (e. g. number, gender, case, person, aspect, tense, ...),
- *link* – syntactic relation[1] between the node and its parent. It can be for example 'предик' (between a verb and its complement), '1-компл' (between a verb and its direct object), 'предл' (between a preposition and a noun), and many others.

The whole corpus contains 32,242 sentences and 461,297 tokens (excluding punctuation). Most of the texts are from journal articles and newspapers, but there are also texts belonging to the fiction genre.

[1] All morphological and syntactic features are described at `http://www.ruscorpora.ru/instruction-syntax.html`.

3 Adaptation of tectogrammatical layer for Russian

Here we discuss the ongoing process of constructing tectogrammatical representation on the basis of morphological information and syntactic relations. The conversion will be described in several steps.

3.1 Format conversion

Both PDT and SynTagRus are represented in XML based formats. In the case of PDT a special PML format was developed [8]. SynTagRus XML format was therefore transfered into PML, so that we can use the TectoMT[1] software framework [12] and TrEd[2] viewer.

As we can see from the corpora description, SynTagRus annotation covers all the features that are necessary to build morphological and analytical layer. The third – tectogrammatical layer will be derived from these two layers in the next steps.

3.2 Converting coordinations

Coordination relations do not belong among dependency relations. Their handling in SynTagRus is different from the PDT style. We will call the coordinated words (or clauses) *coordination members*, the word which governs all the coordination members will be *common parent* and the words depending on all the members will be *common dependents*.

In SynTagRus, according to the Meaning Text Theory [6], the first member of coordination is attached to the common parent. Common dependents are attached to the nearest member, often to the first one. Each other coordination member including conjunctions is attached to the previous member as it is depicted in Figure 3. The edges between coordination members are labeled by 'сочин' (composition relation) or 'соч-союзн' (composition-with-conjunction relation).

In our example, the verbs 'топали' (*stamped*), 'свистели' (*whistled*), and 'расходились' (*left*) are coordinated. They are head of the sentence (the first member is attached to the technical root 'SruA') and have one common dependent, the subject 'Собравшиеся' (*People*), which is attached to the first member 'топали'.

The same sentence but with the coordination handled in the PDT style is depicted in Figure 4. All members of coordination are attached here to the conjunction, the common dependent 'Собравшиеся' is attached also to the conjunction. Members of coordination are distinguished from common dependents with the special attribute '_co'.

The advantages and disadvantages of these two different handling of coordinations are discussed in more detail in [11]. Mel'čuk's approach needs less memory compared to PDT, because it needs no special attributes '_co' for marking coordinating members. It seems that it is also more suitable for annotators (missing '_co' attribute was very common and problematic error in PDT). On the other hand, Mel'čuk's theory

Figure 3: *Handling coordinations in SynTagRus, sentence 'Собравшиеся топали ногами, свистели и нехотя расходились.' (People stamped their feet, whistled and left unwillingly.)*

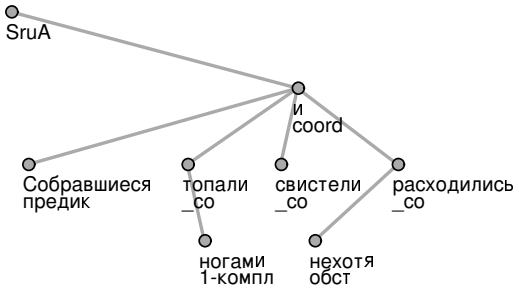

Figure 4: *Handling coordinations in the PDT style (the same sentence as in Figure 3).*

can not reflect inner structure of coordination constructions (for example in *'Peter and Mary or Charlie and Suzanne'*) and does not allow different syntactic relations of coordinated words.

Several problems occurred in automatic conversion of coordinations into PDT style.

Firstly, it is not distinguished in SynTagRus whether a dependent of a member of coordination actually refers to the whole coordination or only to that one member. In our example, the words 'Собравшиеся' (*People*) and 'ногами' (*feet*) are attached both to the first coordination member 'топали' (*stamped*). While 'Собравшиеся' is a common dependent, the word 'ногами' depends on the first member only – on the word 'топали'. The authors of SynTagRus treebank decided not to distinguish them, because this is a notorious source of ambiguity in many cases, for example in *'old men and women'* vs. *'old men and women whose age is not specified'*. Nevertheless, the PDT representation requires this ambiguity to be resolved. The disambiguation can be partially facilitated by a couple of rules. For instance, a subject belonging to the coordinated verbs is almost certain the common subject if there is no other subject in the sentence. This is just the case of the word

[1] `http://ufal.mff.cuni.cz/tectomt`
[2] `http://ufal.mff.cuni.cz/~pajas/tred`

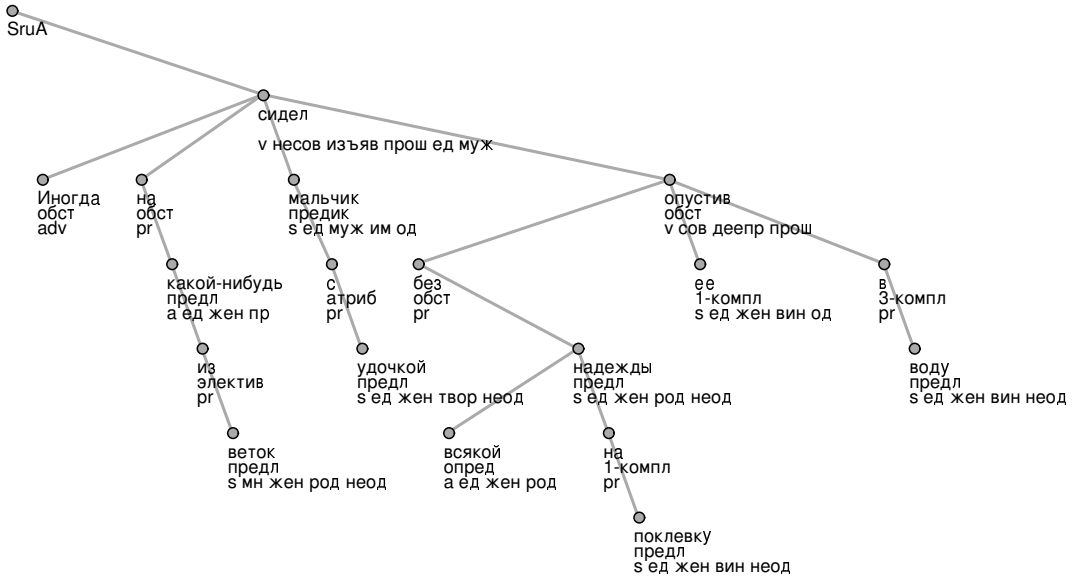

Figure 5: *Analytical representation of the Russian sentence 'Иногда на какой-нибудь из веток сидел мальчик с удочкой, без всякой надежды на поклевку опустив ее в воду. (Now and than a boy with a fishing rode was sitting on a branch, dropping it into the water without any hope to catch fish.)*

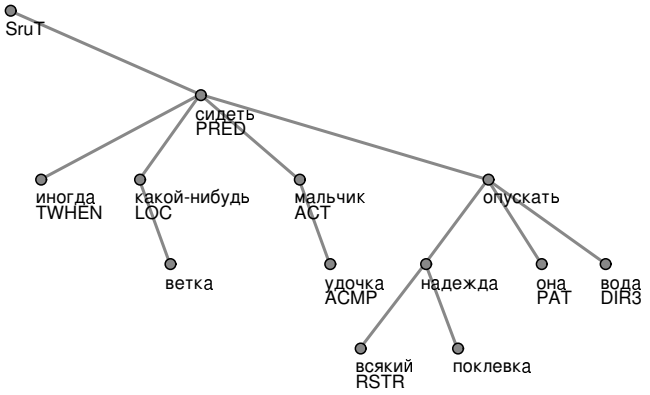

Figure 6: *Tectogrammatical representation of the sentence from the Figure 5, lemmas and functors are depicted.*

'Собравшиеся' (*People*). But by far not all cases can be solved.

Secondly, since punctuation marks are not included in the trees in SynTagRus, it is often the case that there is no node that could serve as the coordination head. In such situations, all coordination members are attached on their common parent instead on a conjunction. We also can not deal with common dependents in such structures, but this problem arises very rarely.

3.3 Function words

Function words (e. g. prepositions, subordinating conjunctions and auxiliary verbs) do not have their own nodes in the tectogrammatical trees. The conversion from analytical trees (in which every word is represented by one node) is done in several steps. Each function word is first marked and assigned to one of the content words. Afterwards the tectogrammatical tree is build using only content (non-function) words as nodes. The meaning of the function words is then expressed by functors and grammatemes (the attributes of respective content-word nodes).

An example of conversion from analytical tree into tectogrammatical tree is depicted in Figures 5 and 6.

For example, the prepositional phrase 'в воду' is represented by the node 'вода' in the tectogrammatical layer. The preposition 'в' is reflected in the functor 'DIR3', which means *to where*.

Some of the rules we use for assignment of function words to content words follow.

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

2. **passive forms** – If there are two verbs which syntactic relation is 'пасс-анал' (analytical-passive) and the lemma of the parent verb is 'быть' (*to be*), the parent verb is assigned as a functional word to the child verb.

3. **future tense** – In Russian (as well as in Czech) future tense of imperfective verbs is expressed analytically as 'to be' + infinitive, e. g. 'будут пользоваться' (*will use*). Therefore, the rule is: If there are two verbs, their relation is 'аналит' (analytical), the lemma of the parent verb is 'быть' (*to be*), and the child verb is in infinitive form, the parent verb is assigned to the child.

4. **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).

5. **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.

3.4 Elided 'to be'

In Czech, personal pronouns in subject positions are often dropped and have to be added (reconstructed) at the tectogrammatical layer. Analogically, we add special nodes into Russian tectogrammatical trees if the Russian verb 'to be' is dropped in the surface sentence shape, as it is for example in 'Я студент' (*I [am] a student*). This is currently approximated by the following simple heuristics: if there is a 'предик' (predicate) relation between two nodes and the parent node is not a verb, then generate a new node labeled with '#ToBe' and attach both previously existing nodes below it (see Figure 7).

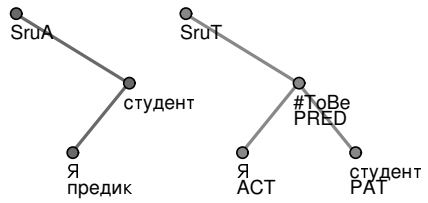

Figure 7: *Adding node '#ToBe' into the tectogrammatical representation of the sentence 'Я студент' (I [am] a student).*

3.5 Assigning functors

Syntactic relations in SynTagRus bare not only syntactic information, but they go deeper towards semantic relations. Labels of semantic relations are called functors in the PDT terminology.

Yet, the classification of this relations within this two frameworks is very different, only a few of them can be mapped as one-to-one. It is for example apposition (syntactic relation 'аппоз' goes to the functor APPS), parenthetical relation ('примыкат' → PAR), and comparative relation ('сравнит' → CPR).

There are five functors for verb arguments in PDT: actor (ACT), patient (PAT), effect (EFF), addressee (ADDR), and origin (ORIG). We expected them to be closely-related to the completive syntactic relations within SynTagRus (1-компл, 2-компл, 3-компл, 4-компл, 5-компл). The apparent correspondence is between the functor PAT and syntactic relation '1-компл' e. g. in 'он читает письмо.PAT' (*he is reading a letter.PAT*). The functor ACT (actor) is often the subject of the sentence and corresponds to 'предик'.

Other relations however do not straightforwardly correspond to the PDT-style functors. In order to assign functors properly we need to know cognitive role of the word, but the argument relations in SynTagRus hardly give this information. Therefore we use several additional rules, for instance: If the relation between a verb and its child node is completive (?-компл) and the child node is a noun in dative case, we assign the functor ADDR (addressee) to it. Example: 'Он дал ребенку.ADDR игрушку' (*He gave to a child.ADDR a toy*).

Some other functors are assigned using lexical list. For example, the words 'чтобы' (*to*), 'в интересах' (*in order to*, 'с целью' (*with the aim of*) usually correspond to the functor AIM. A preposition 'в' (*in*, *to*) corresponds either to the functor LOC (where), if the noun is in locative case, or to the functor DIR3 (to where) for accusative case. A preposition 'в' followed by a noun representing a time, for example *Monday*, *January*, *yesterday*, *week*, corresponds to the functor TWHEN (when). A set of such temporal nouns is not too large to make a satisfactory list of them manually.

You can see an application of the described rules for functors assignment in Figure 6.

4 Small parallel treebank

We have built a small Russian-Czech parallel treebank. Luckily, there exist Czech translations for some of the prose texts included in SynTagRus. We have found one such book which contains Czech translation of one chunk in SynTagRus. We acquired 480 parallel sentences, so that we can compile a small parallel treebank. Whereas the Russian side is largely manually annotated (only the transfer form SynTagRus to tectogrammatics is automatic), the annotation on the Czech side is fully automatic. We use Morce tagger [10], McDonald maximum spanning tree dependency parser [5] and other mainly rule based scripts to generate the tectogrammatical layer. The corpus was compiled using TectoMT [12] framework, which includes all these tools. This parallel treebank, even if very small at the moment, can be once a valuable source of information in comparative language studies.

5 Conclusion and future work

We described the first steps of converting the Russian dependency treebank SynTagRus into the PDT style and developing tectogrammatical layer of Russian. We are on half of the way. We transformed the treebank

into the PDT format, we changed the representation of coordination constructions, because their handling is very different in SynTagRus and in PDT. We hid the auxiliary words, that do not have their own nodes in the tectogrammatical layer, and the elided verbs 'to be' were added. We started with assigning functors (the deep-syntactic relations between tectogrammatical nodes).

In the future, we plan to continue with adding more (often more complex) rules for assigning functors. Other attributes as grammatemes are also going to be assigned to the tectogrammatical nodes.

As for the parsed parallel corpus, we also plan to experiment with aligning the tectogrammatical structures of the two languages on the node level.

Acknowledgments

The work on this project was supported by the grants GAUK 9994/2009, GAČR 201/09/H057, and GAAV ČR 1ET201120505.

We would like to thank Leonid L. Iomdin for providing the complete SynTagRus and for taking the time to answer our questions.

References

[1] S. Cinková, J. Toman, J. Hajič, K. Čermáková, V. Klimeš, L. Mladová, J. Šindlerová, K. Tomšů, and Z. Žabokrtský. Tectogrammatical Annotation of the Wall Street Journal. *Prague Bulletin of Mathematical Linguistics*, (92), 2009.

[2] J. Cuřín, M. Čmejrek, J. Havelka, J. Hajič, V. Kuboň, and Z. Žabokrtský. Prague Czech-English Dependency Treebank, Version 1.0. Linguistics Data Consortium, Catalog No.: LDC2004T25, 2004.

[3] J. Hajič, E. Hajičová, J. Panevová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, and M. Mikulová. Prague Dependency Treebank 2.0. Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.

[4] J. Hajič, O. Smrž, and P. Pajas. Prague Arabic Dependency Treebank 1.0. Linguistics data Consortium, Catalog No,: LDC2004T23, 2004.

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

[6] I. Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, 1988.

[7] J. Nivre, I. Boguslavsky, and L. Iomdin. Parsing the SynTagRus Treebank. In *Proceedings of COLING08*, pages 641–648, 2008.

[8] P. Pajas and J. Štěpánek. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of LREC 2006 Workshop on Merging and Layering Linguistic Information*, Genoa, Italy, 2006.

[9] P. Sgall. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic, 1967.

[10] D. Spoustová, J. Hajič, J. Votrubec, P. Krbec, and P. Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha, 2007.

[11] J. Štěpánek. *Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu*. PhD thesis, Charles University in Prague, 2006.

[12] Z. Žabokrtský, J. Ptáček, and P. Pajas. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on Statistical Machine Translation, ACL*, 2008.