Towards a Latvian Treebank

ABSTRACT

In this paper we describe preparatory work for constructing a Treebank for Latvian as no such resource currently exists. Previously elaborated SemTi-Kamols *hybrid dependency based grammar model has been extended to make it appropriate for broad coverage text annotation. We also have integrated extended* SemTi-Kamols *model with graphical tree editor* TrEd *and complementary toolkit, which originally was developed for Prague Dependency Treebank. Using the obtained environment we have annotated small amount of Latvian text.*

Keywords: Treebank, Latvian, dependency grammar, hybrid grammar, SemTi-Kamols

1. INTRODUCTION

Treebanks are among the crucial resources for the development of NLP tools. For Latvian no such resource currently exists. To address this deficiency the development of Latvian Treebank is ongoing.

As a grammatical framework for the Latvian Treebank, the *SemTi-Kamols* grammar model (Bārzdiņš, Grūzītis, Nešpore, & Saulīte, 2007; Nešpore, Saulīte, Bārzdiņš, & Grūzītis, 2010) is used. It is a hybrid grammar in relation to dependency and phrase structure grammars. This model covers both synthetic and analytical forms of Latvian in a linguistically adequate way. It is not a simple task as Latvian is a highly synthetic language with relatively free word order and rich morphology.

SemTi-Kamols model is strongly based on the pure dependency parsing mechanism described by Covington (2001). Meanwhile it is fundamentally extended with a constituency mechanism to handle analytical multi-word forms consisting of fixed order mandatory words. This enables us to elegantly overcome the limitation of the pure dependency grammars, where all dependants are optional and totally free-order. In *SemTi-Kamols* approach a head and a dependant don't have to be single orthographic words anymore (Bārzdiņš et al., 2007).

Apart from dependency links, the *SemTi-Kamols* model is based on a concept of *x-word*: a syntactic unit describing analytical word forms and relations other than subordination. The concept of *x-word* is analogous in some extent to the Tesnière's *nucleus* — the primitive element of syntactic description introduced by (Tesnière, 1988). From the phrase structure perspective, *x-words* can be viewed as non-terminal symbols, and as such substitute (during the parsing process) all entities forming respective constituents. From the dependency perspective, *x-words* are treated as regular words, i. e., an *x-word* can act as a head for depending words and/or as a dependent of another head word. Similarly as "ordinary" words *x-words* also have rich morpho-syntactic annotation. It is mostly inherited from their constituents, but additional information that specifies the kind of an *x-word* can be included as well, allowing to check for additional agreement restrictions while applying the dependency functions (Grūzītis, 2010).

When integration of *SemTi-Kamols* with *TrEd toolkit* (Hajič, Vidová Hladká, & Pajas, 2001) was started (Pretkalniņa, Nešpore, Levāne-Petrova, Saulīte, 2011), we saw that *SemTi-Kamols* model needs to be extended and clarified to cover texts of different domains and genres. *SemTi-Kamols* in its initial version covers only simple sentences, so the support for composite sentences has to be developed. Also the concept of *x-word* needed to be clarified and developed further.

2. EXTENDED SEMTI-KAMOLS GRAMMAR MODEL

The key question for extending the *SemTi-Kamols* model was the following: what kind of relations do we need to model apart from dependency? The dependency relations in the extended *SemTi-Kamols* model are treated the same way as before. Dependency pairs are the basic relation in the model — they cover subordination by attaching the subordinate element by its governor regardless the position (Nešpore et al., 2010).

The scope of *x-word* was narrowed down by excluding coordination from the *x-word* scope, and one additional construction — punctuation mark construct (*PMC*) — was introduced in the extended *SemTi-Kamols* model. The constructions dealing with other relations than subordination all can be treated similarly as the *x-words* in the initial model: from the dependency view it acts as the regular word, but from the phrase view it act as non-terminal symbol combining its components in the single unit. The distinction among these three constructions is their inner structure — which elements are mandatory, which elements are optional, which elements can act as dependency head and the syntactic relations (or absence of syntactic relations) between the elements.

Thus we arrive at four relation types: dependency, *x-word,* coordination, and punctuation mark construct. Each of these constructions (except coordination, but this may change in future) is divided further in subtypes to give more information about their inner structure and/or functions. *X-words* and coordinated parts of sentence use the rich morpho-syntactic tags developed in the initial grammar model.

2.1. Punctuation Mark Construct

The first relation type introduced anew is punctuation mark construct. The motivation behind this concept is the fact that punctuation in Latvian reflects its grammatical structure. This makes punctuation an essential component to determine the syntactic structure. For example, let us look at two sentences “Sodīt nedrīkst, apžēlot!” and „Sodīt, nedrīkst apžēlot!” („sodīt” — „to punish”, „nedrīkst” — „is not permitted”, „apžēlot” — „to amnesty”). The only difference between these two sentences is the comma, but the first sentence translates as ‘It is not permitted to punish [somebody], [you] must amnesty [him]!’ while the second sentence translates as ‘It is not permitted to amnesty [somebody], [you] must punish [him]!’.

What distinct *PMC* from the phrase-like relations mentioned above (*x-word* and coordination) is its inner structure. *PMC* consists of one mandatory core element, some (usually one or two) optional punctuation mark elements and optional elements which bare no syntactic role in sentence (like addresses, insertions etc.). The mandatory element is the syntactic unit evoking the use of punctuation marks represented by the optional elements. The mandatory element usually is the only *PMC* element which can directly participate in the dependency relation. Elements with no syntactic role usually are *PMC* themselves (see Figure 1) and can have elements participating in dependency relations.

Owing to *PMC* we can handle most of the punctuation usage cases. The most important thing — the clauses of the compound sentence are represented by *PMC* with the predicate as core element (see Figure 3).

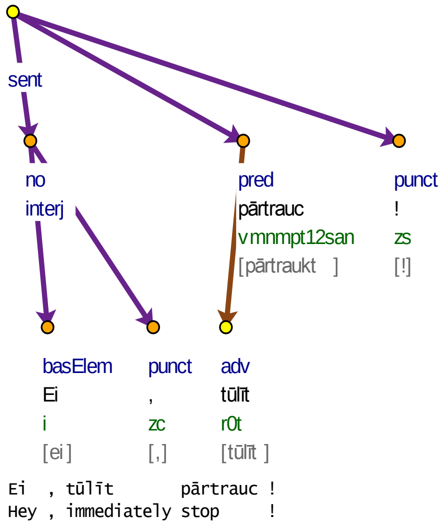


Figure 1. Sentence ‘Hey, stop it immediately!’ demonstrates two *PMC* — overall sentence is a *PMC* consisting from core element ‘*pred*’ and optional elements ‘*no*’ and ‘*punct*’; and element ‘*no*’ also is a *PMC* consisting of an interjection and punctuation

In the initial model only some punctuation was covered and it was done with *x-words*.

2.2. *Coordination*

In the initial *SemTi-Kamols* approach the coordination relation was one of the *x-words*, as coordinated parts of sentence has the same syntactic function in a sentence (Nešpore et al., 2010).

However, the relation between coordinated parts of the sentence is fundamentally different from the relations between the constituents of the analytical forms or multi-word units, therefore in the extended model the coordination was distinguished as a separate relation. This brings the *SemTi-Kamols* model even closer to the Tesnière's structural syntax, where coordination (*jonction*) is one of the basic concepts. Coordination (horizontal) relationship differs from a subordination (vertical) relationship, it is formed by two or more homogenous nodes that have the same function but these nodes are not constituents of one nucleus like multiword units (Tesnière, 1988).

The coordination relation can link different types of syntactic units, therefore in the extended model the same relation is used to represent both coordinated parts of sentence (see Figure 2) and coordinated clauses (see Figure 3). If it links coordinated parts of sentence, it is annotated with morpho-syntactic tag inherited from those coordinated parts.

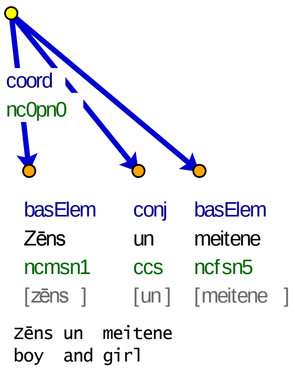


Figure 2. Fragment 'The boy and the girl' demonstrates the coordinated parts of sentence

Elements composing coordination structure can be divided in two types — elements representing coordinated parts and supporting elements (conjunctions and punctuation marks). Coordination structure must consist of at least two coordinated parts and usually at least one supporting element between each two coordinated parts. Only coordinated parts can act as heads of dependency.

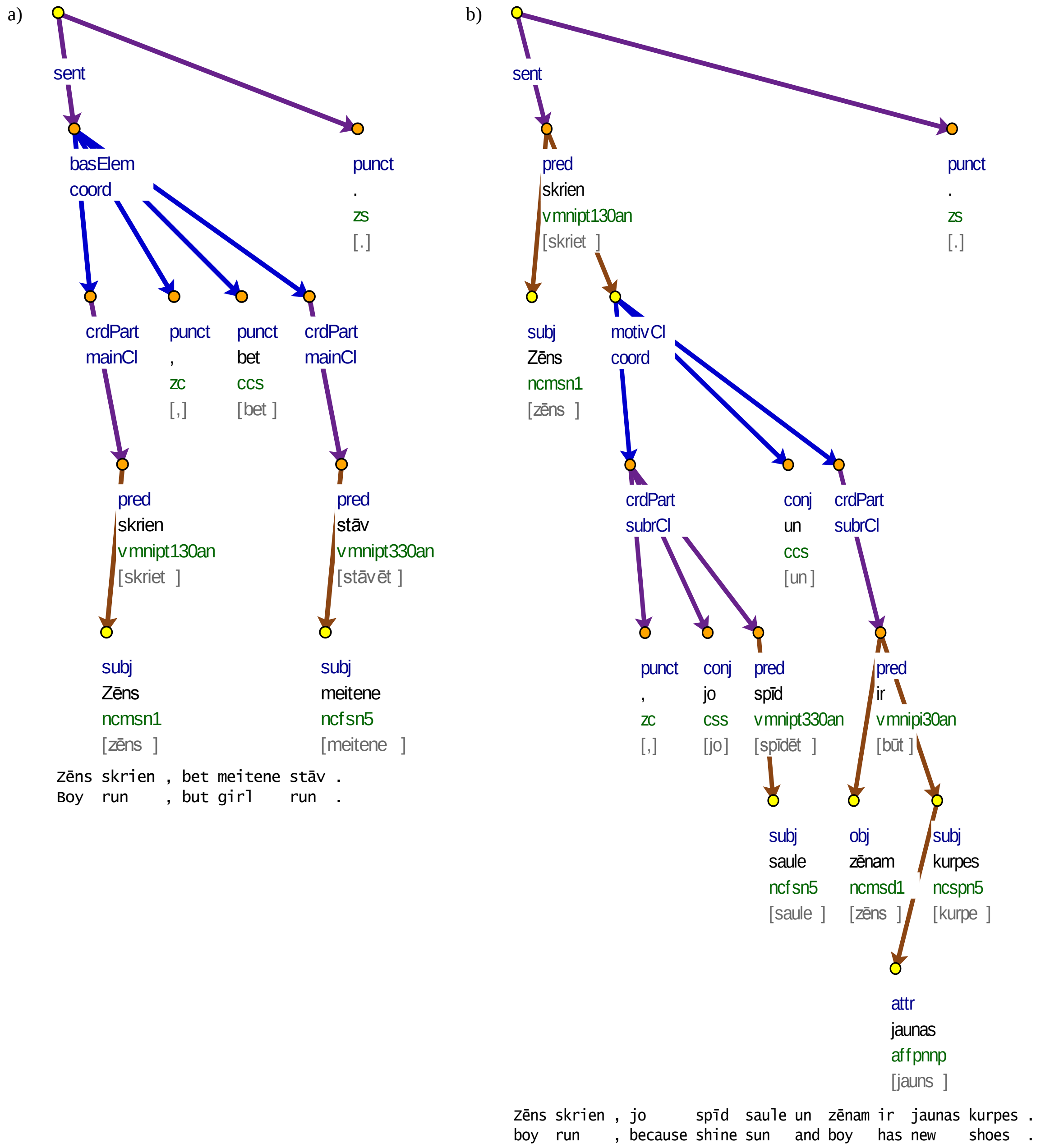


Figure 3. Sentence ‘The boy is running, but the girl is standing.’ (a) demonstrates coordinated main clauses. Sentence ‘The boy is running because sun is shining and the boy has new shoes.’ (b) demonstrates coordinated subordinate clauses.

2.3. *X-word*

X-word in the extended *SemTi-Kamols* somehow comes back to its original concept — being a multiword unit where every element is mandatory.

X-words are used to describe various syntactic constructions, though relations between elements in the inner structure of *x-words* are different. This information is reflected indirectly by the type of the particular *x-word* (*x-Verb*, *x-Preposition*, *x-Apposition*, etc.). This type also determines how the morpho-syntactic tag for the *x-word* is obtained and which *x-word* elements can act as dependency heads.

We have following types of *x-words* for Latvian. First are *analytical forms:* perfect tenses of verb (*x-Verb*, see Figure 4 a) and prepositional phrases (*x-Preposition*, see Figure 4 b). *X-Verbs* and *x-Prepositions* are formed by one content word and one or several function words. *X-Preposition* combines a preposition (rarely postposition) and a noun (or a pronoun), *x-Verb* combines at least one auxiliary verb and one content word (participle, noun, adjective, adverb or pronoun) (Nešpore et al., 2010). In these constructions usually only content word can act as dependency head.

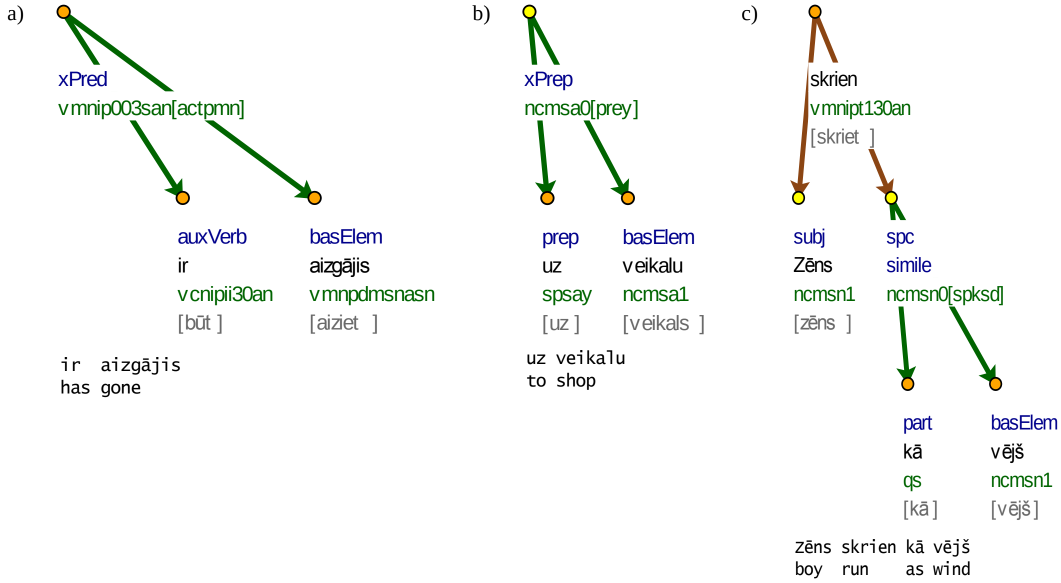
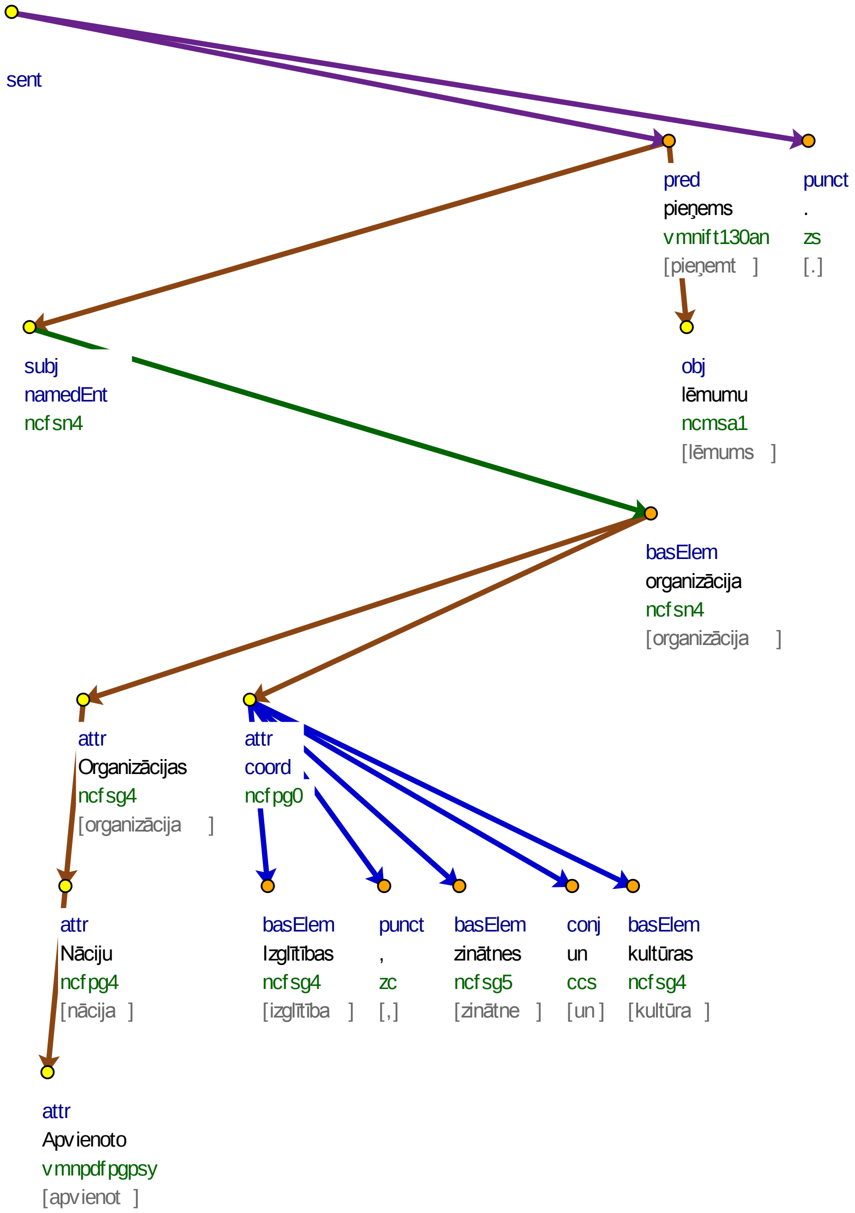


Figure 4. Fragment '[he] has gone' (a) demonstrates *x-Predicate*; fragment 'to shop' (b) demonstrates *x-Preposition*; fragment 'The boy runs like a wind' demonstrates simile and the way how *x-words* are incorporated in the syntax tree.

Second type of *x-word* is *simile* (see Figure 4 c). It is formed by one content word and one function word. In this case also only content word can act as a dependency head.

Third type of *x-word* is multiword units (*named entities*, *analogues of wordgroup*, *idioms* and *multiword numerals* and *appositions*). The distinctive feature of this type of *x-words* is that no element of these *x-words* can be used as dependency head, thus all the elements of these *x-words* will occur in the text one right after another.

Annotating *named entities* and *idioms* is one of easiest sources to the ambiguous annotation of the Treebank — distinguishing whether the given fragment of a text is an idiom or not often relays on an annotator's previous experience and subjective interpretation. When the annotation is done by multiple annotators, it is easy to obtain different annotations to the same text strings. This was the main concern why we decided to annotate inner syntactic structure of the idioms and the named entities that have clear tree representation (see Figure 5). In this way the representation of a string as an idiom or named entity becomes more similar to the case when the same string is not recognised as idiom or named entity, thus making post-processing of such potentially ambiguous mark-up easier.



```
Apvienoto Nāciju  Organizācijas Izglītības , zinātnes un  kultūras organizācija pieņems   lēmumu   .
united    nations organization  education  , science  and culture  organization will_take decision .
```

Figure 5. Sentence ‘United Nations Educational, Scientific and Cultural Organisation will make the decision.’ demonstrates how *named entity* is used; everything in the subtree below ‘*namedEnt*’ belongs to this *named entity*

3. INTEGRATION WITH TREd TOOLKIT

We have integrated the extended *SemTi-Kamols* model with *TrEd toolkit* — tools developed for Prague Dependency Treebank (Hajič, Böhmová, Hajičová, & Vidová Hladká, 2000). The central tool in this toolkit is *TrEd* — customisable graphical editor for tree-like structures. The default data format for *TrEd* toolkit is Prague Markup Language (*PML*) (Pajas, & Štěpánek, 2006). It is *XML* based mark-up language developed to suit the needs of the linguistic annotations. It is independent from annotation scheme; it supports multilayer annotations and offers verification by the *PML schema*. *TrEd toolkit* also includes tools for querying treebanks and tools for batch processing trees (Hajič et al., 2001).

We have developed *PML* profile for extended *SemTi-Kamols* model annotations, thus obtaining *XML* based data format for Latvian Treebank (Pretkalniņa et al., 2011). Also we have developed an extension module for *TrEd* to enable full *TrEd* support for our format (Pretkalniņa et al., 2011). The extension we developed contains stylesheets, *PML* schemas for our data format and macros to automate common annotation tasks.

Using all the above mentioned *TrEd* can be used as an environment for manual creating/editing Latvian Treebank.

4. Summary

Preparatory work for Latvian Treebank development is successfully ongoing. We have extended *SemTi-Kamols* dependency based hybrid grammar model to fit most syntax constructions of Latvian by additional relations — like *punctuation mark construct* — and clarifying the existing relations — like *x-words* and coordination.

We have developed extension module enabling us to use graphical tree editor *TrEd* as an annotation environment.

Using the obtained results we have created small Treebank as a proof of concept. We have annotated first 100 sentences of J. Gaarder's "Sophie's World" (Pretkalniņa et al., 2011) and ~100 sentences of Latvian fiction text.

Even the annotated text amount is still small, it contains the broad coverage of syntax constructions of Latvian, and thus we estimate that *SemTi-Kamols* model is very close to cover all Latvian.

For creating bigger Treebank we are working on integrating the obtained environment with *SemTi-Kamols* rule-based partial parser (Bārzdiņš et al., 2007).

Rerferences

Bārzdiņš, G., Grūzītis, N., Nešpore, G., & Saulīte, .B. (2007). Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*, (pp. 13–20).

Covington, M.A. (2001). A Fundamental Algorithm for Dependency Parsing. In *Proceedings of the 39th Annual ACM Southeast Conference*, (pp. 95–102).

Grūzītis, N. (2010). *Formal Grammar and Semantics of Controlled Latvian Language*. Summary of Doctoral Thesis in Computer Science. Riga, University of Latvia.

Hajič, J., Böhmová, A., Hajičová, E., &Vidová Hladká, B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, Amsterdam: Kluwer, (pp. 103–127).

Hajič, J., Vidová Hladká, B., & Pajas, P. (2001). The Prague Dependency Treebank: Annotation Structure and Support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia, USA, (pp. 105–114).

Nešpore, G., Saulīte, B., Bārzdiņš, G., & Grūzītis, N. (2010). Comparison of the SemTi-Kamols and Tesnière's Dependency Grammars. In *Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective*, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, (pp. 233–240).

Pajas, P., & Štěpánek, J. (2006). XML-Based Represen-tation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, (pp. 40–47).

Pretkalniņa, L., Nešpore, G., Levāne-Petrova, K., & Saulīte, B. (2011). A Prague Markup Language Profile for the SemTi-Kamols Grammar Model. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA)*, (pp. 303–306).

Tesnière, L. (1988). *Основы структурного синтаксиса.* (Trans.) Ред. В.Г. Гак. Москва: Прогресс (Original work published 1959).