L. Pretkalniņa, K. Levāne-Petrova

PREPARATORY WORK FOR LATVIAN TREEBANK[1]

Abstract. In this paper the preparatory work for Latvian Treebank has been examined. We use the SemTi-Kamols dependency based hybrid grammar model integrated with the TrEd toolkit originally developed for Prague Dependency Treebank. We have notably extended and adapted both the SemTi-Kamols model and TrEd to fit them to our needs. As the result the small (~200 sentences) manually annotated Treebank covering typical syntax constructions of Latvian has been created.

1. Introduction

Treebanks are among the crucial resources for the development of NLP tools. For Latvian, a highly synthetic Baltic language with relatively free word order, no such a resource currently exists. Latvian Treebank project has been recently launched to address this deficiency.

Latvian is a highly inflective language where synthetic forms prevail, but also analytical forms are present, and thus we have chosen to model it trough a hybrid approach. In essence we are trying to model synthetic forms with the help of dependencies, but analytic forms – with the help of phrase structures.

[1] This work is funded by the State Research Programme "National Identity" (project No 3) and the Latvian Council of Sciences project No 09.1544. "Application of Factored Methods in English-Latvian Statistical Machine Translation System".

2. SemTi-Kamols

We have chosen to use the SemTi-Kamols[1][2] grammar model as a syntactical framework for Latvian Treebank. In essence, SemTi-Kamols model is close to the Tesnière's dependency grammar[3]. To make it appropriate for broad coverage text annotation, SemTi-Kamols model has been notably extended, by elaborating the treatment of horizontal relations. The extended model has four types of relations: vertical – dependency, and horizontal – coordination, x-words, punctuation mark constructs (PMC).

Dependency is the basic relation in the SemTi-Kamols model. It attaches the subordinate element with particular morphological features by the governor regardless its position in the sentence.

In the SemTi-Kamols model, a **horizontal relation** can be analysed from two viewpoints. From the dependency view it act as a regular word, but from the phrase view it act as a non-terminal symbol combining its components in a single unit that can further participate either in some other phrase-like structure or in a dependency relation as head or dependant. Elements of horizontal relations can participate in dependencies as heads but not as dependants. Elements of horizontal relations can also be x-words, PMCs or coordination themselves.

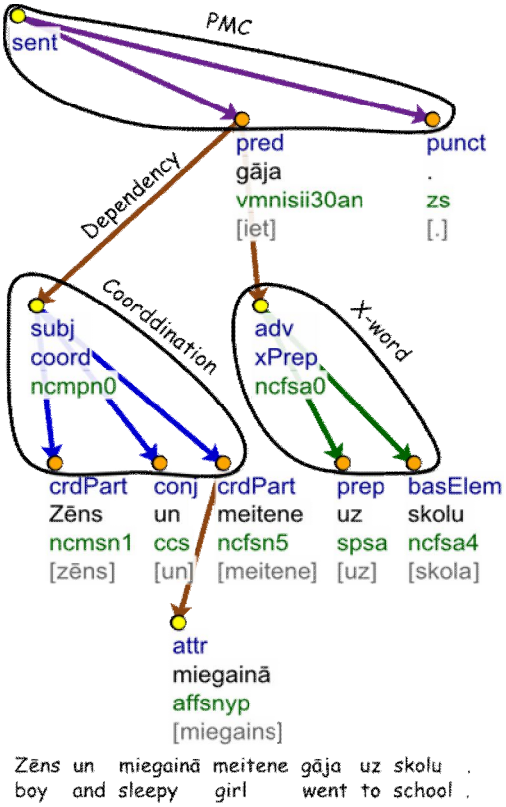
---

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

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

[3] *Tesnière L.* Éléments de syntaxe structurale. Klincksieck, Paris, 1959. (Translation to Russian: *Теньер Л.* Основы структурного синтаксиса/ Ред. В.Г. Гак. Москва: Прогресс, 1988.)

An **x-word** (see figure) is a structure consisting of several elements, all of which are mandatory. The x-word performs a single syntactic function in the sentence.



Sample sentence «The boy and the sleepy girl went to the school.»

X-words cover analytic forms, similes and multiword units – named entities, idioms, multiword numerals, analogues of subordinate wordgroups. X-words are divided further in subtypes – each of above

mentioned phenomena has its own subtype. Subtypes further specify which of x-word constituents can act as a dependency head. Usually it is one specific element of an x-word (e.g., if x-word is prepositional construction, then preposition can not act as a head), but for some x-words, like named entities, no elements will ever invoke dependencies.

Each x-word has a morpho-syntactic tag – similar to those morphological tags that regular words have. This tag specifies the role that an x-word can fulfil in the sentence and the type and/or inner structure (implicitly). Having this tag ensures the unified way parsing both regular words and x-words.

The **coordination** relation in the SemTi-Kamols model corresponds to the Tesnière's concept of *junctions*.

In the SemTi-Kamols model, coordination (see fig.) is handled in the phrase-structure style – both coordinated elements and conjunctions and/or punctuation between coordinated elements are united to be treated as a single unit in further analysis. Only coordinated elements can act as dependency heads.

SemTi-Kamols uses the same coordination relation to handle both coordinated parts of sentence and coordinated clauses. In the case of coordinated parts of sentence, coordination construction is supplemented with the mofo-syntactic tag inherited from its constituents – similarly as with x-words.

Punctuation mark construct (PMC) is a relation introduced to handle punctuation mark usage in Latvian. The motivation behind this concept is the fact that punctuation in Latvian reflects its grammatical structure, thus making it essential for syntax analysis.

PMC (see fig.) is a phrase-like structure which consists of one mandatory core element, some (usually one or two) optional punctuation mark elements and optional elements that bare no syntactic role in the 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.

PMCs are divided further in subtypes (e.g., direct speech, insertion) reflecting the motivation behind the use of punctuation. With the help of PMC the sentence's structure in terms of clauses is annotated.

3. Integration with the TrEd toolkit

The TrEd[1] toolkit was used to develop the Prague Dependency Treebank[2]. Its central tool TrEd is a visual tree editor that offers support for both dependency and constituency trees. The toolkit also contains tools for batch processing and querying treebanks.

We have integrated the SemTi-Kamols model with the TrEd toolkit by developing a profile of Prague Markup Language (PML) and a TrEd extension module working with this data format. In this way we have enabled the hybrid tree support in TrEd. PML usage also allows us to take advantages of the XML-based data format.

As a proof of concept, we have annotated ~100 sentences of text originally in Latvian and the first 100 sentences of J. Gaarder's "Sophie's World", in lines with the project of Parallel treebank of North European languages[3].

---

[1] *Hajič J., Vidová Hladká B., Pajas P.* The Prague Dependency Treebank: Annotation Structure and Support // Proceedings of the IRCS Workshop on Linguistic Databases. Philadelphia, 2001. P. 105–114.

[2] *Hajič J., Böhmová A., Hajičová E., Vidová Hladká B*. The Prague Dependency Treebank: A Three-Level Annotation Scenario // Treebanks: Building and Using Parsed Corpora. Amsterdam: Kluwer, 2000. P. 103–127.

[3] The Sofie Treebank – A Parallel Treebank of North European languages. **www.hf.uio.no/iln/om/organisasjon/tekstlab/prosjekter/arkiv/sofie.html**

4. Conclusion

The development of the Latvian Treebank is successfully ongoing. We have obtained an appropriate grammar model and integrated it with the leading TrEd toolkit thus obtaining a convenient environment for manually creating the treebank. Our future plans involve integrating TrEd with the SemTi-Kamols parser[1] to enable semi-automated annotation process.

Currently the SemTi-Kamols grammar model has been tested only on Latvian, however, it is developed in a way it can be applied for other languages as well, especially synthetic inflective ones, like Baltic and Slavic.

The SemTi-Kamols' basic concepts – the hybrid dependency and phrase structure grammar, punctuation mark handling, etc. are language independent. The language specific part in the SemTi-Kamols model is the subtypes of x-words and PMC, thus adapting this model for other languages can be achieved through deciding what language contractions will be modelled in which way and what subtypes for each or relations are needed.

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