Syntactic Issues Identified Developing the Latvian Treebank

Lauma PRETKALNIŅA[1] and Laura RITUMA
Institute of Mathematics and Computer Science, University of Latvia

Abstract. The Latvian Treebank is being developed since 2010. In this paper we describe the latest developments of this project and the problems currently faced. We examine several gaps in our annotation scheme like determinant, ellipsis and insertion annotation and describe solutions we have chosen.

Keywords. Latvian Treebank, SemTi-Kamols, dependency grammar, hybrid grammar, annotation guidelines, determinant, ellipsis, insertion, parenthesis, attribute.

Introduction

We have been working on Latvian Treebank since 2010 [1]. The treebank currently contains approximately 1500 manually annotated sentences from various genres of text.

The Latvian Treebank utilizes an extended SemTi-Kamols grammar model [2]. It is a hybrid grammar in relation to dependency and phrase structure grammars. We consider four distinct relation types in the grammar model [1]: dependency, x-word, coordination, and punctuation mark construct (PMC), illustrated in Figure 1. The basic and most commonly used relation in the model is the dependency, used to model the subordination relations in the sentence. X-word is a phrase structure that covers analytical word forms and relations other than subordination and coordination (for example, named entities, prepositional constructions, multiword numerals etc.). Coordination is a relation between two or more syntactic units with the same syntactic function in the sentence. Coordination is used to represent both coordinated parts of sentence and coordinated clauses. PMC is the way to link the punctuation mark to the syntactic structure. This is important as the punctuation in Latvian reflects the grammatical structure.

Since the latest report [1], the scope of linguistic phenomena covered by this grammar model has been significantly extended. A new syntactic role — determinant — has been introduced to describe linguistic phenomena not covered by the initial model. Subtypes of x-word and coordination constructions have been clarified (for example, introducing a specific subtype to describe coordination parts with a generalizing word). Also, a methodology for handling particles has been introduced.

However, during the annotation of new texts we have identified several gaps in our grammar model that we describe in detail in the following chapters:

[1] Corresponding author: lauma@ailab.lv.

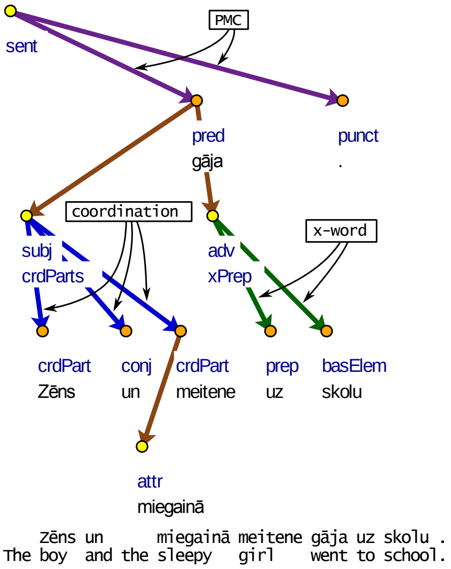

Figure 1. Latvian Treebank annotation example

- It is unclear how to annotate the part of the sentence when one syntactic relation includes more than one semantic relation (for example, different attribute types, determinants);
- The model does not currently describe attachment rules for grammatical structures that, unlike most parts of sentence, might also refer to the whole sentence and not only a single part of sentence (situants, determinants, etc).
- There is no consensus on how much information should be restored in case of ellipsis ('reduction' in Latvian linguistic tradition).

A vast majority of these problems arise in the *gray zones* of traditional Latvian syntax theories [3]. Although the basis of Latvian syntax theory has been laid in the beginning of 20th century, later it has been strongly influenced by Russian and Czech linguistic theories (formalism and structuralism). Later modern linguistic theories have complemented our theory, focusing on the functions and semantics of syntactic constructions. As a result, the syntax theories of the phrase[2], parts of the sentence, simple sentence are carefully developed, but the next layers of syntax (composite sentences and text) are inadequately studied and described [4]. Research is continued using modern linguistic theories, however there are still a number of language phenomena that are not revised in new perspective or it is done incompletely. When facing phenomena not covered by current Latvian theories, we must seek our own solutions. When defining annotation guidelines, we want to keep our grammar model maximally informative and consistent. However, we also must keep our grammar model simple enough for annotators and end-users, both human and software tools.

Considering the Russian and Czech influence on traditional Latvian syntax theory, we are comparing our annotation principles to Prague Dependency Treebank (PDT)

[2] In this paper we shall use "phrase" to such a word-group where the components are connected with subordination relationship.

annotating principles [5]. Unlike our Treebank, PDT uses a pure dependency grammar, but there are similarities in both models and they can be compared. For comparison, we also consider more complicated syntactic annotations in the Russian National Corpus [6], which also uses dependency based annotation with more syntactic relation types than in PDT [7].

In this paper we will describe some problematic cases in Latvian, which arise in different layers of syntax (the syntax of phrase, sentence or text).

1. Attributes and attributive relation

In Latvian traditional syntax 'attribute' is defined as a part of the sentence dependent of a noun. Attribute usually is positioned before a noun and expressed by an adjective, a numeral, a declinable participle or a noun in genitive. [8]

However, when it comes to characterizing deverbal nouns, the situation becomes more complicated. E.g. In sentence *Man ir sapnis* <u>*par savu māju*</u> 'I have a dream <u>about [my] own house</u>' noun *sapnis* 'dream' is characterized by prepositional construction *par māju* 'about the house', so we could say that the prepositional construction is an attribute. From the semantic point of view, this construction is similar to a object relation determined by the verb valence, e.g. *Ansis sapņo* <u>*par savu māju*</u> *'*Ansis is dreaming <u>about [his] own house'</u>, where *par māju* is the object of verb *sapņot* 'to dream'. The phrase *sapnis par māju* 'a dream about the house' contains two semantic relations — attributive and object relation. We have to consider if we want to annotate these constructions as a separate specific type of relation or ignore them, describing only the formal syntactic attributive relation.

We can compare the previous example with *Man patīk tā māja* <u>*mežā*</u> 'I like that house <u>in the woods'</u>. The noun 'mežā' (*in the woods*) with adverbial meaning characterizes the noun 'māja' (*house*), not the verb *patīk* 'like', but the construction *māja mežā* 'house in the woods' is not considered as phrase in Latvian traditional syntax. The problematic part of sentence is positioned after the noun it characterizes, which is not common for attributes in Latvian.

For comparison, in PDT such attribute-like parts of sentence with subject or object meaning are annotated as attributes. Attribute-like parts of sentence with adverbial meaning are annotated as borderline cases with a special role AtrAvd or AdvAtr to show the ambiguity of these constructions [5].

For Latvian Treebank we consider following possible solutions:

1. Annotate all attribute-like constructions as attributes. This approach is the simplest, but also the less informative.
2. Introduce a finite set of attribute variations to reflect all above described differences. This approach is the most informative, but it needs additional research about possible attribute variations.

Choosing from both options, we must take into account that there are syntactic relations refer to the whole sentence, not to a specific word and phrase, and we are annotating them separately. For example, *sapnis par māju* 'a dream about the house' is considered as phrase, but not *māja mežā* 'the house in the woods'. Therefore currently in Latvian Treebank the parts of the sentence that have both attributive and adverbial meaning are annotated as semi-predicative components, as they hold the information about the place or the time of someone's or something's existence (the existence is considered as implicated secondary predicate in the current sentence). Other attribute-like members of sentence are annotated as attributes, no further distinction is made.

2. Identifying and annotating determinants

Determinant is defined [8] as a free part of the sentence, which refers to the whole sentence and is not related with any specific part of a sentence. (It means that this syntactic relation is fulfilled only in sentence, not in the phrase.) Usually determinants are placed in the beginning of sentence. There are two kinds of determinant distinguished — determinant with adverbial meaning (in Latvian tradition it is called 'situant', in world's linguistic it is close to understanding of 'sentence adverb' [9]), e.g. *Pļavā skrien zirgi un rāpo gliemeži* 'In the meadow horses run and slugs crawl' and determinant with syncretic subject and object meaning (experiencer, possessor, beneficiary), which usually is expressed by noun or pronoun in dative, e.g. *Man ir vīrs un divi bērni* 'I_{dative} have a husband and two children'.

For Treebank purposes we need clear guidelines for both identifying and annotating determinants. While the subject/object determinant is quite easy to identify in most situations due to its dative case and meaning, the identification of the situant can be quite ambiguous in cases when word order has been changed due to the communicative structure of the sentence. In these cases non-valent adverbial modifiers should be considered as situants, but all others — as adverbial modifiers subordinated to predicate. However, practical application of this principle is complicated because for some modifiers it is hard to unambiguously define if they are valent or not, and the development of the first valence lexicon for Latvian has just started [10].

In the PDT's annotation subject's dative is mentioned — it is a type of free dative, who is not determined by verb or adjective [5]. It is consistent with our understanding of determinant as a free part of sentence. In Prague Dependency Treebank free subject's dative is annotated as an object, but determinants with adverbial meaning (situant) are not annotated as different members of sentence.

For Latvian Treebank we are treating determinant and situant separately from adverbial modifiers and objects because they form a specific syntactic relation that, unlike other members of sentence, can apply to the whole sentence (also to more than one clause). We have considered following possible solutions.

1. Annotate with special determinant/situant roles only determinants relating to two or more coordinate clauses, e.g. subject/object determinant — *Man salst rokas un dreb kājas* 'My hands are freezing and [my] legs are shaking'; situant — *Pļavā skrien zirgi un rāpo gliemeži* 'In the meadow horses run and slugs crawl', but in other cases annotate them as objects or adverbial modifiers. The advantage of this approach is that lots of identification ambiguities are eliminated. The main disadvantage is inconsistent annotation between simple sentences and composite sentences, e.g. sentences *Pļavā skrien zirgi un rāpo gliemeži* and *Pļavā skrien zirgi* 'In the meadow horses run' are annotated differently: *pļavā* in the first sentence would be annotated as a situant, but in the second — as an adverbial modifier. The other disadvantage is that the specific information about determinant relation (e.g. about syncretic subject/object determinant) is lost in unannotated cases.

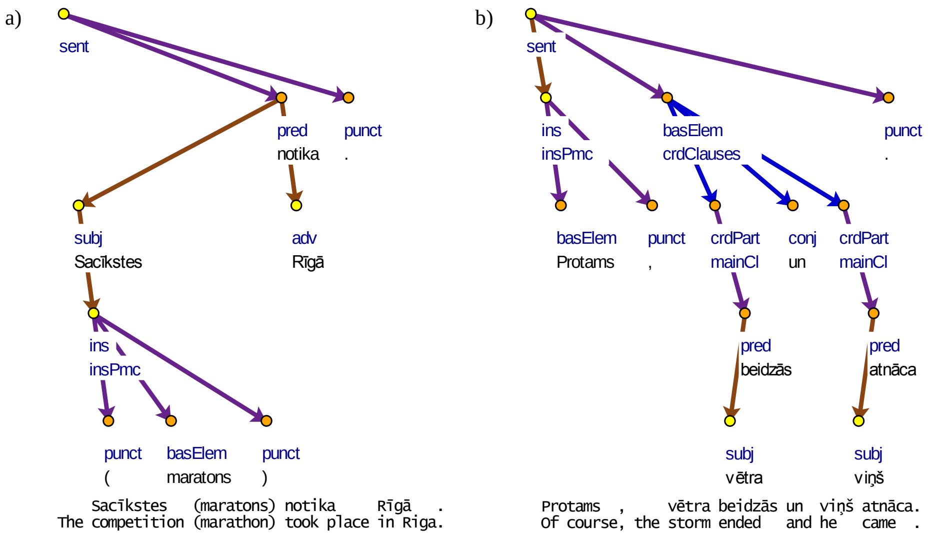

Figure 2. Insertion annotation in Latvian Treebank

2. Annotate determinants also in simple sentences, taking into account the position in the sentence for determinants with adverbial meaning (determinant must be placed in the beginning of the sentence) and only the syncretic object/subject meaning for determinants in dative. The advantage of this approach is that such an annotation is more informative. The main disadvantage is that we would annotate as a situant also a valent adverbial modifier of a verb that is placed in the beginning of the sentence because of the actualization.
3. Annotate determinants considering their meaning and relation with other sentence members regardless their position in the sentence. It means to annotate the adverbial determinant as a situant in all cases, when it is not related with valence of the verb. It would be the most informative solution, but it is not possible until an extensive lexicon of valences is developed.

Currently we annotate the determinants in both composite and simple sentences like it is described in the second solution. To identify a determinant/situant in simple sentences we will only look at the following features — meaning, free syntactic relation with sentence and position in the sentence (usually a determinant/situant is placed in the beginning of the sentence, but not always). Also we hope that in future a lexicon of valences will help to solve the problem of determining situants/determinants in simple sentences.

3. Syntactic treatment of insertions and parenthesis

In the traditional Latvian syntax insertions and parenthesis are defined as syntactically independent units which are not members of the sentence and have epistemic or evidentially modal meaning (insertion) or have explanatory or clarifying meaning (parenthesis) [8]. In practice, these units often feature something from both meanings, both insertion and parenthesis can have very different syntactic forms, and their unclear syntactic relations with sentence do not allow us consistently distinguish insertion from parenthesis [11]. So in further text we will use term ‘insertion’ to describe both insertions and parenthesis.

Insertion itself can be in different forms — wordform, semi-predicative component or predicative clause. Furthermore, insertion can be related to either a single member of sentence or, just as determinants, to a clause or several coordinated clauses [11]. Linguists of other languages have already studied these syntactic constructions and partly defined syntactic relations in different cases, but so far these studies and their findings are not integrated in current Latvian syntax theory and we lack more research directly on Latvian syntactic constructions of insertions.

Still, in our Treebank we want to include as much information as possible, so we want to show directly which unit is related with insertion, even if we cannot determine the type of the syntactic relation.

For comparison, in PDT the term ‘parenthesis’ is used, and it is concerned as an additional adjunction of a remark to the statement included in the sentence. The speaker usually uses parenthesis to explain something, to add some remarks, to express his/her emotions, to apologize, to refer to something, etc. The necessary condition to annotate a construction as a parenthesis is graphic separation marks. Otherwise parenthesis is considered as a member of the sentence. If removing the punctuation would result in valid sentence structure, then such parenthesis can be annotated with the standard role with an extra tag added to specify the parenthesis function (for example Adv-Pa). If the parenthesis is predicative unit and does not fit syntactically in the structure of the sentence, it is suspended to predicate of the sentence and gets the role of parenthesis. The same solution is used if the parenthesis doesn’t look like predicative unit, but also doesn’t fit in the structure of the sentence. Unlike us, in PDT identity forms and abbreviations in the brackets are not considered a parenthesis, but an apposition. However, all occurrences of vocatives are assigned as parenthesis in the PDT [5].

In the current annotation guidelines we link the insertions through the dependency link to the related unit. This results to a similar representation to determinants if insertions are related to clauses (see Figure 2b) and similar representation to members of sentence if insertions are related to a member of sentence (see Figure 2a). In both cases ‘insertion’ role is used for the dependency link. When an insertion refers to the whole sentence, we choose not to attach them to the predicate, but to the root of tree in both simple and composite sentences for consistence reasons, like determinants mentioned before.

4. Ellipsis

While the problems described in previous chapters mostly arise from gaps in Latvian syntax theory, the decisions related to the ellipsis annotation are more technical. To achieve a more precise depiction of the syntactic tree of the sentence, the omitted elements can be represented with accordingly annotated empty node in the tree. It is possible to annotate the new node either with exact wordform or with morphological pattern showing the features that are uniquely defined by context. Still, we need precise guidelines how to decide which of the omitted elements should be represented as artificial nodes in the sentence tree.

PDT utilizes a more simplified approach — in case of ellipsis, if omitted element has a dependent, then the dependent takes the place of the omitted element in the tree and is annotated with a special role, identifying the fact of ellipsis. They do not annotate ellipsis if the omitted element is: 1) a copula in the predicate with a nominal part, 2) an adjective (sometimes), 3) subject, 4) governing clause between noun and adverb, 5) counted units. In these cases dependent gets the role of a reduced element [5].

In Russian National Corpus a more complicated solution has been chosen, that is more similar to our method. The principles of annotating ellipsis are the following: 1) if omitted element is found in another part of sentence, it is restored with an exact lemma and wordform, 2) if the omitted verb is not mentioned in the sentence before and we cannot determine the exact lemma and wordform, an artificial word is inserted (like an artificial node in our case). Reduced units are restored to show the full structure of the sentence [12], [13].

At the beginning we considered to restore the omitted elements if they are heads of dependency. If it was possible to determine the exact unit form context and situation, we showed the exact lemma and wordform. The problem is that the inter annotator agreement is very low in such cases. Practice shows that structure restored from context and situation is highly subjective, and looking for the exact unit in previous sentences (context) to is very time consuming.

To reduce the amount of manual annotation work and ambiguities, we decided to put the following restrictions on ellipsis annotation:

1. Ellipsis is annotated only with information contained in the current utterance or sentence. No information from context outside the current sentence is added — it means that in future we will not include the information of exact wordform if it is not clear from current utterance, even if we could find that information in other sentences.
2. Omitted copulas and omitted auxiliary or modifier are annotated as ellipsis. This is done to reflect the full information about structure of predicate, as this information reflects the morphosyntactic agreement between parts of sentence and may be important for the development of data driven syntax parsers in future.
3. Any other omitted element is restored if it is inner node of the tree (i.e., it is a head of some dependency and has a nonempty ancestor).

5. Conclusions

The traditional Latvian syntax includes a lot of semantic features, and it is not always possible to define precisely the phenomena that should be shown in a purely syntactic annotation, as in the earlier examples of different attributes or determinants. In addition, the annotation of ellipsis shows that it is quite difficult to determine exact meaning or even structure of an omitted part of the sentence because of the ambiguity of the language.

To simplify the annotation process, we must draw a clear border between different layers of syntax (sentence and text) and between semantic roles and syntactic roles. To show the information omitted in the current annotation system, we are considering to develop an additional annotation layer similar to the tectogramatical layer in PDT. It could solve some of the abovementioned problems: in the next layer we could show syncretic semantic relations of attribute-like elements, but in current layer we could leave them as attributes. In the next layer we also could show the exact reduced lexeme if it appears in the context sentences or can be identified from the situation, but in current layer we could show only the fact that there is a reduced element in the sentence structure.

In dealing with the problem of the border between determinants and dependent parts of sentence we hope to use the valence lexicon that is now in development.

6. Acknowledgements

This work is supported by National Research Programme “National Identity (language, Latvian history, culture and human security)” and European Regional Development Fund under the project No. 2011/0009/2DP/2.1.1.1.0/10/APIA/VIAA/011.

References

[1] L. Pretkalniņa, G. Nešpore, K. Levāne-Petrova, B. Saulīte, *Towards a Latvian Treebank.* Actas del 3 Congreso Internacional de Lingüística de Corpus. Tecnologias de la Información y las Comunicaciones: Presente y Futuro en el Análisis de Corpus, eds. Candel Mora M.Á., Carrió Pastor M., 2011, pp. 119–127.

[2] G. Nešpore, B. Saulīte, G. Bārzdiņš, N. Grūzītis, *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, Vol. 219, IOS Press, 2010, pp. 233–240.

[3] L. Ceplītis, J. Rozenbergs, J. Valdmanis, *Latviešu valodas sintakse.* Zvaigzne, Riga, 1989.

[4] J. Rozenbergs, *Par latviešu valodas sintakses zinātnes attīstību*. Available online: http://www.liis.lv/latval/teksts/rozenb1.htm

[5] E. Hajičová, Z. Kirschner, P. Sgall, *Manual for Analytical Layer Annotation of the Prague Dependency Treebank (English translation)*. Charles University, Prague, 1999. Available online: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html

[6] Homepage of Russian National Corpus, available online: http://www.ruscorpora.ru/en/index.html, Institute of Russian language, Russian Academy of Sciences, 2003–2012.

[7] И.М. Богуславский, Н.В. Григорьев, Л.Л. Иомдин et al., *Разработка синтаксически размеченного корпуса русского языка*. Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». University of Saint Petersburg, Saint Petersburg, 2002, pp. 40–50.

[8] V. Skujiņa et al.,*Valodniecības pamatterminu skaidrojošā vārdnīca*. LU LVI, Riga, 2007.

[9] P.H. Matthews, *Concise Dictionary of Linguistics*. Oxford University Press, Oxford, 2007, pp. 364.

[10] G. Nešpore, B. Saulīte, N. Grūzītis, G. Garkāje, *Towards a Latvian Valency Lexicon*. Proceedings of the 5th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, IOS Press, 2012, to be published.

[11] L. Rituma, *Iespraudumafunkcijas teikumā un tekstā*. University of Latvia, Riga, 2011.

[12] Ю. Д. Апресян, И. М. Богуславский, Б. Л. Иомдин et al., Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы, *Национальный корпус русского языка: 2003–2005*, ed.М. Индрик, 2005, 193–214.

[13] I.M. Boguslavsky, S.A. Grigorieva, N.V. Grigoriev, et al., *Dependency Treebank for Russian: Concepts, Tools, Types of Information*, Proceedings of the 18th Conference on Computational Linguistics, Vol 2, Saarbrücken, 2000, pp. 987–991.