Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-107

Lithuanian Dependency Treebank ALKSNIS

Agnė BIELINSKIENĖ [a], Loïc BOIZOU [a], Jolanta KOVALEVSKAITĖ [a] and Erika RIMKUTĖ [a,1]
[a] *Vytautas Magnus University, Centre of Computational Linguistics, Lithuania*

Abstract. This paper aims to describe the on-going work on creation of the Lithuanian syntactically annotated corpus ALKSNIS focusing on its structure, morphological and syntactic annotation principles. The corpus is scheduled to be completed at the end of 2016, and it should reach about 2350 sentences from texts of various genres. ALKSNIS is based on a dependency model. The corpus is provided in two formats: PML (Prague Markup Language), as a core format, and PAULA XML. The compilation of the list of abbreviations for syntactic labels and collecting of the information about the presentation of the syntactic relations and dependences were based on the experience (with some changes) of Czech researchers [1]. At present, 18 main syntactic labels (excluding variants) are used in ALKSNIS.

Keywords. Lithuanian language, treebank, morphological annotations, syntactic dependencies

1. Introduction

In 2015–2016 at the Centre of Computational Linguistics (CCL) at Vytautas Magnus University (VMU) a syntactically annotated Lithuanian corpus ALKSNIS is being created; it is planned to be the gold standard of syntactic analysis. The corpus started with the establishment of the national consortium CLARIN-LT in Lithuania as well as with the beginning of the implementation of the project *Lithuanian Membership in International Scientific Research Infrastructure – Common Language Resources and Technology Infrastructure Consortium* (CLARIN ERIC[2]). ALKSNIS will be available via the CLARIN infrastructure and will be prepared following the standards employed within this infrastructure.

The first attempts to prepare an experimental Lithuanian Treebank were made in 2007–2008 in CCL at VMU during the project *Internet resources: Annotated Corpus of the Lithuanian Language and Tools of Annotation (ALKA 2)*. The annotated texts are taken from the newspaper domain and thus represent the normative Lithuanian language. The treebank contained 1,566 sentences and 24,265 tokens. This treebank was designed without a proper standard and was considered to be poorly designed to provide a useful basis for such a fundamental resource. The syntactic annotation scheme only distinguishes 5 basic grammatical relations (subject, object, predicate, attribute and

[1] Corresponding Author, Vytautas Magnus University, Centre of Computational Linguistics, K. Donelaičio str. 52-206, LT-44244 Kaunas, Lithuania; E-mail: erika.rimkute@vdu.lt.
[2] http://clarin.eu/

modifier) plus an additional underspecified relation for other dependencies between words and a special relation for words attached to an (implicit) artificial root node. This corpus was used as a training corpus for statistical dependency parsing [2]. In 2013, another attempt was made by the Institute of Lithuanian Language [3][3].

In this context the new treebank ALKSNIS was started from scratch, using the parser ANTIS [4] created in 2014. The corpus is scheduled to be completed at the end of 2016, and it should reach about 2350 sentences from texts of various genres (see Section 2). One of the main aspects of an on-going project is the preparation of annotation guidelines. ALKSNIS is mainly inspired by the Lithuanian grammar tradition, which is a loosely described dependency model, and the Czech experience in formalizing language as represented by the morphosyntactic layer of Prague dependency Treebank[4].

2. Composition and Format

The corpus ALKSNIS will consist of several text types: newspapers, journals, fiction and legal texts (see Figure 1).

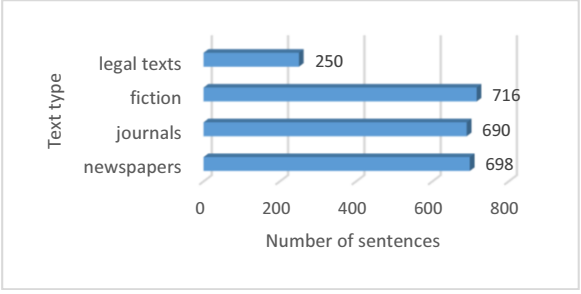

Figure 1. Composition of ALKSNIS.

In comparison to treebanks of other languages (Swedish Treebank[5] or Prague Dependency Treebank), with its 2,354 sentences (approx. 35,000 tokens) ALKSNIS is a rather small treebank. When the comparison is made with the other small language from the same Baltic group, Latvian (3,882 sentences and 53,225 tokens) [5], we can infer that the size of the Lithuanian Treebank can be taken as quite reasonable, especially, as a starting point. In the future, ALKSNIS has to be developed, and new data (also from spoken language) will be added.

Many other treebanks usually contain newspaper texts from various domains. Aiming at broadening the spectrum of genres, we adopt the same practice as used in such treebanks as Prague Dependency Treebank, Swedish Treebank and Dependency Treebank for Russian [6], and include not only newspaper texts, but, also, fiction and legal texts into ALKSNIS. Newspaper texts are complete articles from Lithuanian newspapers and journals. The fiction part is built by collecting full texts of small genres of prose (short stories, essays); legal texts are represented with such genres as orders and regulations. All texts are published in the period of 2004–2014.

[3] http://www.lki.lt/LKI_LT/index.php?option=com_content&view=article&id=803&Itemid=153
[4] https://ufal.mff.cuni.cz/pdt3.0
[5] http://stp.lingfil.uu.se/~nivre/swedish_treebank/

ALKSNIS is provided in two formats: PML[6] (Prague Markup Language), as a core format, and PAULA XML. PML is the native format of TrED (developed by Charles University in Prague[7]), which is particularly well designed for tree visualisation and redaction. However, in order to provide a convenient interface with extended search possibilities, a tool was designed to convert data from PML format to Paula XML format. The latter format is then converted into the native ANNIS format [7] using a conversion tool Pepper (developed by Humboldt University in Berlin). It allows ALKSNIS to be available online using the ANNIS server (Humboldt University in Berlin)[8] (see 4.1). The choice of ANNIS as a visualisation and query tool, instead of PML-TQ, allows us to complement other CCL morphologically annotated data with ALKSNIS data.

3. Annotation Process

The experience from other languages shows that automatic tools speed up the annotation process: the correction is faster than full human redaction of the whole data. Consequently, all texts integrated in the corpus are firstly annotated by automatic means. The process of word and sentence segmentation and morphological analysis is carried out by the tools provided by the web service semantika.lt[9]. The results are presented in the JSON format. Then, these results are used by the rule-based syntactic analyser ANTIS to generate dependency analyses of each sentence, and the results are given in the PML format. All the sentences are manually checked by one linguist and corrected by a group of linguists. It is obvious that it would be better if the same text was annotated by two or three linguists, and then their results of annotation would be automatically compared. At present, each linguists annotates separately, as there is not much time given for the preparation of the corpus, neither there are enough human resources. Besides, as long as the annotation guidelines are not completed, during the discussion of the annotation results within the linguist group, annotation guidelines are being changed.

The labels and the guidelines for the Lithuanian syntactic annotation are prepared following the Czech experience, as the systems of both languages are similar [1]. However, there are some differences discussed in chapter 4.2.

4. Annotation Levels

In PML files, each node of a tree corresponds to a word, a punctuation mark or the other text element (symbol, digit, etc.) within a sentence. The following information is presented for each node: 1) a used form; 2) a lemma; 3) a morphological tag, and 4) a syntactic function (subject, object, etc.). Dependencies are shown by links between words (see Figure 2).

The visualisation of the same sentence “Taip pat jau rezervuota pusė ploto kitais metais iškilsiančiame statinyje” (*Also, half of the area has already been reserved for a construction to be built next year*) by PML and PAULA versions:

[6] http://ufal.mff.cuni.cz/jazz/PML/index_en.html
[7] https://ufal.mff.cuni.cz/tred/
[8] https://zenodo.org/record/20713#.V3z8x9aw-ic
[9] http://www.semantika.lt/TextAnnotation/Annotation/Annotate

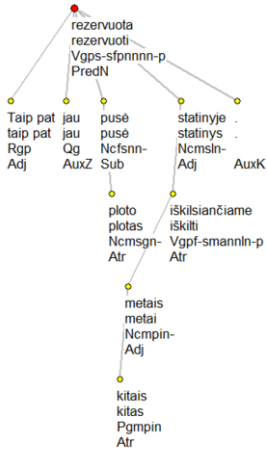

Figure 2. PML version.

In PAULA files, there are the following levels: a token, a sentence, a part of speech, morphological features, a syntactic function, and dependency links. The part of speech and morphological features are conflated in one field in PML files (see Figure 3).

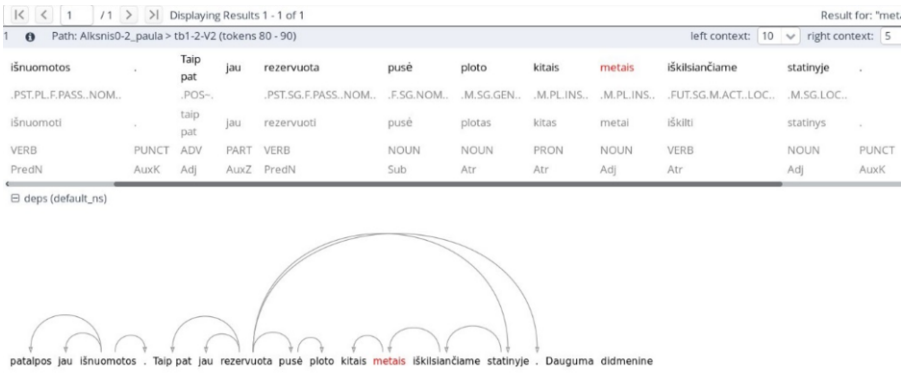

Figure 3. PAULA version.

4.1. Morphological annotation

In a tagger traditional Lithuanian grammatical categories are used. Some semantic features are added (person names, geographical names). We consider abbreviations and shortenings as parts of speech. As mentioned above, we use two different file formats: PML and PAULA. Further, we describe the main differences between these formats. Although the list of morphological categories and values is the same for both the PML and PAULA versions of ALKSNIS, their surface representations differ significantly.

4.1.1. PML

In PML files, morphological annotation is provided according to a morphological standard inspired by MULTEXT-East[10] format. Each word is associated to a string in which each character expresses a morphological value, e.g., Vgmp3s--n--ni- (verb, general, main form, present tense, 3rd person, singular, no gender, no voice, not negative, no definiteness, no case, not reflexive, indicative mood, no degree), for the word *turi* ('has'). The first character in capital letters indicates the part of speech. The number of morphological categories, that is, the length of the string, depends on parts of speech (from 2 to 14). A dash indicates that the feature is irrelevant for the annotated word (see Figure 2).

Adapted MULTEXT-East format used for morphological annotation in PML file is convenient for the sake of briefness, but it makes the search more difficult. Indeed, it is structured as a code vector where the value of each code depends on its position and on a given part of speech. For example, the code *a*, depending on the part of speech and the position in the code vector, may indicate the accusative case, the simple past or the active voice.

4.1.2. PAULA

To avoid the previously mentioned problem with complicated search options feature names are converted during the PAULA XML format generation, so that each morphological feature gets an unambiguous name. Feature names are taken from Leipzig glossing system[11], because these names (e.g., *ACC* for the accusative case) are shorter than the universal dependency features, where both the feature type and the feature value are indicated (e.g., *Case=Acc*). There are several additions to the Leipzig glossing system for lacking features. These additions to the standard list are prefixed by a tilde, e.g., *~ACT* for the active voice.

In order to minimize the number of nodes, which has a direct influence on the efficiency of the ANNIS system, morphological information of each word is provided as a list of features, and the search for feature values is done using regular expressions inside the morphological field. All features are surrounded by dots, e.g., .F1.F2.F3., in order to avoid ambiguities. For example, *M*, the identifier for masculine gender, matches a part of the nominative case tag, *NOM*. Consequently, searching for **M** in the morphological field would provide inaccurate results. The dots allow an unambiguous search for **.M.** instead (see Figure 3). As mentioned, in PAULA files, the parts of speech are provided in a specific POS field distinct from the morphological field. The POS identifiers are taken from the Universal dependency POS tagset[12].

4.2. Syntactic annotation

ALKSNIS is based on a dependency model. Such models proved to be suitable for a typologically similar language, e.g., Slavonic (PDT, SynTagRus[13]), with a relatively free word order and rich inflection. While the Latvian treebank uses a hybrid model with a dependency core extended by some constituency extensions, ALKSNIS relies on a more traditional dependency framework.

[10] http://nl.ijs.si/ME/V4/msd/html/index.html
[11] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf
[12] http://universaldependencies.org/u/pos/index.html
[13] http://www.ruscorpora.ru/en/

The syntactic annotation is the same in both PAULA and PML format: dependency relations (marked as edges/curves between lexical nodes) + a common tagset inspired by the Prague Dependency Treebank. Syntactic dependencies are presented according to the hierarchy starting with the main sentence component (usually a verb or a conjunction mean), which is the root of a sentence. The other dependent components are combined by edges/curves with their main components creating syntactic dependences.

4.2.1. Syntactic tags

In AKSNIS the following syntactic functions are defined: a predicate, a subject, an object, an attribute and modifiers. The functions are indicated with such labels: respectively, *Pred, (PredN, PredV), Sub, Obj, Atr* and *Adj*, as well as others denoting, for instance, the way sentences are joined. A complete list of labels with examples is provided in the Table 1. At present, 18 main syntactic labels (not including variants) are employed. In some cases, there is a need for double syntactic functions, thus, double labels are used too (e.g., *Pred_Atr*). In case of coordination, the coordinated components or predicates of sentence components are indicated in a syntactic label by an under-dash and a label *Co*, e.g., *Pred_Co*, which means that two predicates are joined by a coordination relation. Double labels are also used when subordinate clauses are annotated; then it is necessary to indicate the type of a clause, i.e., such a label is used for the predicate of a subordinate component, e.g., *Pred_Adj*.

Like it was mentioned before, during the annotation, we relied on the Czech experience. However, there are some differences in the Lithuanian treebank: some syntactic labels differ, and the understanding of grammatical relations, too. First of all, the structure of the tree is different. In the Lithuanian treebank, the top of the tree starts directly with the syntactic root of a sentence, while the Czech researchers have introduced an intermediate position or even a separate label *AuxS*, which is a starting point to draw dependencies. We have decided not to use part of the Czech labels, e.g., apposition (*Apos*), which in ALKSNIS gets the same label like an attribute, as well as various types of Aux, for example, co-referential pronouns (*AuxO*), emphasizing words (*AuxZ*) (in ALKSINIS *AuxZ* signifies the function of a particle) and others. Such ambiguous labels as Czech *AtrObj* or *AtrAdv* and others are not marked either, because we are trying to create rules that help to distinguish between unclear and ambiguous cases (we discuss such cases with prepositions for location or direction (see 4.2.2), etc.).

The Czech researchers make a distinction between labelling of a compound analytical predicate whose auxiliary "to be" is marked by *AuxV* and compound nominal and verbal predicates, the labels of which are respectively different. We do not make such a distinction, but we separate nominal (*PredN*) and verbal (*PredV*) predicate components, and we always label an auxiliary word or a conjunction simply as Pred.: *…bus (Pred) atliktas (PredN) ir (yra) sunku (PredN) tikėtis (PredV) (…will be done and it is difficult to expect)*. The label *PredN* is used not only for the cases when there is a copula "be" (the Czechs use *Pnom*), but, also, for other conjunctions, e.g., *jaučiuosi (Pred) lieknas (PredN) (I feel slim)*.

It has been decided to mark ellipsis only where a part of a sentence is clearly omitted and, for this reason, it is impossible to determine a syntactic relation. Ellipsis is marked when a predicate is omitted (in such cases a dash or nothing is used); also, when a word performs the function of the omitted part of a sentence (e.g., *Matėme raudoną (Obj_ExD), kuris skrido (We saw red which was flying)*. We do not mark ellipsis in nominative and

comparative sentences, if a copula is omitted, etc. The Czech treebank and grammars have a wider understanding of ellipsis.

Table 1. Syntactic functions and abbreviations in ALKSNIS.

Abbreviation	Syntactic function	Example
Sub	Subject	Jis (Sub) sakė (*He said*)
Pred	Predicate (or auxiliary word)	Jis ėjo (Pred) (*He was going*)
PredN	Nominal part of compound nominal predicate	Buvo (Pred) patenkintas (PredN) (*He was pleased*)
PredV	Verbal part of compound verbal predicate	Turi atsilyginti (PredV) (*You have to repay)*
Obj	Object	Laukiu svečio (Obj) (*I am waiting for a guest*)
Atr	Attribute (or apposition)	Vidurinė (Atr) mokykla (*a secondary school*)
Adj	Adjunct	Partizanų gatvėje (Adj) (*in Partizanų street*)
Aux	Auxiliary function (e.g., a comma or other symbols)	
AuxC	Subordinated conjunction	Sakė, kad (AuxC)… (*He/she/they said that…*)
AuxK	Terminal punctuation of sentence	
AuxL	Auxiliary lexical unit (e.g., a foreign word)	JAV kompanija North (AuxL) American (AuxL) Investment (AuxL) Consulting (AuxL) Inc. (Atr) (*a US company North American Investment Consulting Inc.*)
AuxP	Preposition	Pereis į (AuxP) lygį (*He/she/they will move to the level*)
AuxZ	Particle	Kaip ir (AuxZ) visada (*as always*)
Coord	Coordination node	Gamintojų ir (Coord) pardavėjų (*producers and sellers, Gen.*)
_Co (e.g. Sub_Co)	Coordinated words (e.g. coordinated subjects)	Gamintojų (Atr_Co) ir (Coord) pardavėjų (Atr_Co) (*producers and sellers, Gen.*)
Par	Parenthesis	Tiesą sakant (Par), atrodo… (*To tell the truth, it seems…*)
ExD	Ellipsis or other omissions in a sentence	Moteris – (Pred_ExD) būtybė (*a woman – a being*)
Pred_ (e.g. Pred_Obj)	Type of a subordinate clause (indicated next to a predicate)	Rašė, kad ypatybes lemia (Pred_Obj)… (*He/she/they wrote that the features depend on…*)

4.2.2. Annotation difficulties

Although the annotation guidelines are increasingly stable, they are still in progress, so it is difficult to make a final assessment. The process of the corpus preparation is still continuing, therefore, there are a few issues that have to be solved. Firstly, some uncertainties of syntactic annotation need to be clarified, because it is not always possible to determine a syntactic function without criteria set beforehand. In Lithuanian, it is difficult to distinguish between a location and an object when they are expressed with prepositions. It is still not clear how to treat clarifications: should they be considered as modifiers or should they be labelled as coordination. In Lithuanian dictionaries, quite a

lot of problems are caused by interpretations of parts of speech, especially for form-words.

5. Future Work

Information enrichment. One direction for future development is related to an explicit annotation of MWEs in ALKSNIS. In Lithuanian, there are several types of MWEs [8]: nominal (named entities, idioms and collocations), verbal (idioms, collocations), proverbs and MWEs of grammatical nature, e.g., multi-word adverbs, multi-word prepositions. At present in ALKSNIS, we annotate all words of different MWE types separately, except for those of grammatical nature, which are treated as single lexical units already on the morphological level, and appear as single nodes in the tree structures. For future applications, it would be useful to annotate all words of different MWE types separately (except for those of grammatical nature), but to have a special annotation level "MWE" for all MWEs.

The other direction would be thematic role annotation. ALKSNIS already has an experimental layer with thematic role annotation. This information is not yet provided externally and even not corrected by linguists. Both extensions will be considered as the next steps.

Format extension. The mapping of ALKSNIS into the universal framework is not included into the current starting stage of the treebank, because the treebank is developed as a very light project. Nevertheless, the authors are aware that this is a significant trend. The choice of UD part of speech categories in the ANNIS version is the first step in this direction. Since by the end of 2016 we have to prepare the corpus of 2350 sentences and lack human resources, we are short of time to transform the corpus into the UD format.

New resources. When the syntactic annotation of the corpus is complete, we are planning to create a statistically based parser by employing the syntactically annotated corpus ALKSNIS as the gold standard.

References

[1] J.Hajič, J.Panevová, E.Buráňová, Z.Urešová, A.Bémová, *Annotations at Analytical Level. Instructions for Annotators* (11.10.1999), UK MFF ÚFAL Praha, 1999.

[2] J.Kapočiūtė-Dzikienė, J.Nivre, A.Krupavičius, Lithuanian Dependency Parsing with Rich Morphological Features, *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages* (2013), 12–21.

[3] D.Šveikauskienė, Lietuvių kalbos sintaksinė analizė, *Lietuvių kalba* **7** (2013), 1–20.

[4] L.Boizou, F.Zamblera, Syntactic Engine for the Lithuanian Language, *Proceedings of the Sixth International Conference Baltic HLT 2014* (2014), 69–75.

[5] L.Pretkalniņa, L.Rituma, Construction in Latvian Treebank: the Impact of Annotation Decisions on the Dependency Parsing Performance, *Proceedings of the Sixth International Conference Baltic HLT 2014* (2014), 219–226.

[6] I.Boguslavsky, I.Chardin, S.Grigorieva, N.Grigoriev, L.Iomdin, L.Kreidlin, N.Frid, Development of a Dependency Treebank for Russian and its Possible Applications in NLP, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)* (2002), 852–856.

[7] T.Krause, A.Zeldes, ANNIS3: A new architecture for generic corpus query and visualization, *Digital Scholarship in the Humanities* **31(1)** (2016), 118–139.

[8] J.Kovalevskaitė, E.Rimkutė, L.Boizou, Representation of MWEs in the Lithuanian Dependency Treebank, *The 6th PARSEME general meeting*, 7-8 April 2016, Struga, FYR Macedonia.