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

doi:10.3233/978-1-61499-701-6-136

Universal Dependency Treebank for Latvian: A Pilot

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

Abstract. In this paper we present the first Universal Dependency Treebank for Latvian. Latvian UD Treebank contains approx. 1 thousand sentences. It has been created from Latvian Treebank newswire texts with the help of an automatic conversion. This resource is an important prerequisite for integrating Latvian in various international language processing frameworks and making Latvian data more welcoming to international researchers. This paper also includes an analysis of the main conversion problems and describes known discrepancies between annotations in Latvian UD Treebank and Universal Dependency annotation guidelines.

Keywords. Universal Dependencies, Latvian, Treebank, syntax

1. Introduction

In this paper we present the first treebank of Latvian annotated according to the Universal Dependency (UD) [1] Treebank guidelines[2]. Universal Dependency initiative aims to provide a universal set of categories and guidelines for syntactically annotated corpora to facilitate creation of consistently annotated resources across multiple languages [2]. Creating a UD treebank is an important prerequisite for integrating Latvian in various international language processing frameworks, and making data that has already been annotated for Latvian more welcoming to international researchers.

Latvian UD Treebank is created from newswire texts (a part of Latvian Treebank [3][4]) with the help of an automatic conversion procedure. Latvian UD Treebank has been published together with the Universal Dependency Treebank version 1.3, which includes 54 treebanks and represents 40 languages. It contains 1K sentences with approx. 20K tokens. By joining Universal Dependencies, Latvian UD Treebank has been made available to anyone with a Creative Commons license BY-NC-SA. For example data from Latvian UD Treebank has been used in training [5] Google's SyntaxNet based parser[3].

[1] Corresponding Author, Email: lauma@ailab.lv
[2] Universal Dependencies online documentation http://universaldependencies.org
[3] Experiment description in Google Research Blog https://research.googleblog.com/2016/08/meet-parseys-cousins-syntax-for-40.html;
results from https://github.com/tensorflow/models/blob/master/syntaxnet/universal.md

2. Data

The data used in the creation of Latvian UD Treebank was taken from Latvian Treebank. Latvian Treebank (created by University of Latvia) is currently the only publicly known syntactically annotated corpus for Latvian. It consists of manually created syntax annotations and automatically created, human checked POS tags. The treebank is natively annotated accordingly to a hybrid in-house grammar model, which models the core structure as a dependency tree augmented with some phrase-like constructions. It contains 3.8 thousand sentences and 53 thousand tokens.

Morphological features of a token in Latvian UD Treebank are represented with a positional tag. The meaning of each character position in the tag depends on the part of speech [6]. Currently there are 13 parts of speech used. Each token has 1 to 11 features depending on its part of speech.

Latvian Treebank's native annotation model is a dependency based hybrid. In this model each sentence is modeled as a tree, where dependency links connect either single words or multi-word phrases. Both full phrases and parts of phrases (single words) can act as dependency heads in this model [7][3].

There are three types of phrases distinguished in Latvian Treebank: x-word, punctuation mark constructs (PMC) and coordination. X-word is a phrase consisting of several words fulfilling a single syntactic slot, e.g., perfect tenses, prepositional constructions, named entities, multiword units, etc. A PMC is a "phrase" consisting of one or more punctuation marks and the base word invoking the use of these punctuation marks, e.g., *Smagi strādājot, viņa nogura* /hard work.CVB she tire.PST3/ 'while working hard she got tired' — comma and *strādājot* forms a PMC. A coordination "phrase" consists of conjuncts, conjunctions and punctuation marks used for separating conjuncts.

Latvian Treebank also includes ellipsis annotation in a following fashion: if there is a node missing from the tree, but the node's dependents are present, then a special reduction node is inserted in the tree and an appropriate morphological tag is assigned.

For now, the Latvian UD Treebank only contains the newswire data of Latvian Treebank, however in future versions it is planned to expand Latvian UD Treebank to cover other parts of Latvian Treebank as well.

3. Conversion Procedure

To obtain a Universal Dependency treebank conforming all possible guidelines of annotation an elaborate conversion procedure was created. The conversion consists of three deterministic steps — retokenization, obtaining morphological information and obtaining syntactic annotations.

3.1. Tokenization

In most cases tokenization in Latvian Treebank corresponds to what is expected for UD Treebanks, however, Latvian Treebank annotates complex conjunctions and particles as "words with spaces", e.g. *lai gan* 'even though'. For UD these words are transformed as constructions of several tokens, linked with the *mwe* relation (first token as the head and the rest — as its dependents), see Figure 6. Morphological tags for these constructions are assigned as follows — last token gets the tag from the original

"word with spaces" and other tokens are annotated as particles. Review for these situations was done, concluding that for data currently in Latvian UD Treebank this heuristic gives 100% correct result. Similarly, in most cases lemmas from Latvian Treebank can be directly used in UD Treebank. In the case of a "word containing spaces" splitting the lemma at the point of the white space gives correct results in all of the cases observed in the newswire part of Latvian Treebank.

Currently Latvian UD Treebank features a discrepancy against UD guidelines: reflexive verbs are not split into two tokens — verb and reflexive pronoun, as in Latvian the reflexive pronouns are deeply infused into verbs in such cases, and the reflexive marker 's' featured in most verb forms traditionally is considered to be a part of the ending, not absorbed by the word [8]. This final mostly reflexive verb forms (*mazgāties* 'to wash [oneself]' vs *mazgāt* 'to wash', *mazgājos* '[I] wash myself' vs *mazgāju* '[I] wash'), however, both past participle forms (reflexive *mazgājies* '[he] has been washing [oneself]' vs non-reflexive *mazgājis* '[he] has washed') end with 's'.

3.2. Morphological annotation

The morphological annotation of words in UD represents the information about a lemma, part-of-speech and lexico-grammatical features. Most of this information can be obtained directly from morphological tagging of the Latvian Treebank. POS categories in Latvian Treebank and UD are largely similar with a few exceptions.

- For some pronouns to distinguish pronouns (*PRON*) from determiners (*DET*), syntactic labeling must be used — pronouns in the attribute role are tagged as determiners according to UD guidelines, as formally and morphologically they are indistinguishable. For example, *tā* /that.F.SG.NOM/ can be either demonstrative pronoun as in *tā meitene* 'that girl', or determiner as in *tā ir meitene* 'it is a girl'.
- A subordinated clause in Latvian Treebank can be linked to its head by words tagged as subordinating conjunction, relative pronouns, some adverbs or, rarely, nouns in prepositional constructions. All these words should be annotated as SCONJ according to UD guidelines. However, the current annotation scheme does not allow to identify such adverbs and nouns as opposed to other adverbs or nouns.
- A notable source of ambiguous annotation are words tagged as residuals and abbreviations in Latvian Treebank. UD guidelines require annotating them as common/proper nouns, adjectives, adverbs etc. wherever it is possible, however there is not enough information to make correct distinction. This may lead to incomplete/incorrect annotations on the syntax level, too (see Section 3.3). Currently we employ heuristics to distinguish certain residual classes — abbreviations consisting of capital letters only (*NATO*) are tagged as PROPN, abbreviations like *k-dze* 'Mrs.' are tagged as NOUN, and abbreviations such as *u.c.* 'etc.' are tagged as SYM.
- Some adverbs and adjectives derived from participles do not have verbal annotations in Latvian Treebank, thus it is not possible to add an appropriate inflectional UD features *VerbForm* (*Part*, *Trans*) and *Voice* (*Act*, *Pass*), e.g., word form *nepieciešams* 'necessary' lacks features *Part* and *Pass* in Latvian UD Treebank, as it has been annotated as an adjective in the original data.

While assigning morphological features, the most interesting decisions had to be made regarding participles and other verbal derivatives. In Latvian there are four types of participles, two types of non-inflected converbs and one type of partially inflected converb. In lines with original annotations and lemmas these words were assigned *VERB* POS. As participles behave like adjectives and feature the same lexico-grammatical properties, they were assigned *VerbForm=Part*. Converbs usually express a second predication and behave differently than any other POS in Latvian, thus, the decision of which *VerbForm* to assign was hard. Currently we use *Trans* for both non-inflected and partially inflected converbs, even though all adverbials in Latvian are non-inflected.

Doing the conversion another interesting difference between Latvian traditional grammar and UD guidelines surfaced — in Latvian only verb moods inflected for person and number are considered to be finite while in UD every verb form except infinitives and participles/gerunds/transgressives are considered finite. This influences how certain grammatical moods are analyzed — debitive/necessative (*jāskrien* 'must run'), conditional (*skrietu* '[someone wishes for he/she/it/they] was/were running' or '[if he/she/it/they] was/were running') and quotative (*skrienot* '[someone said, he/she/it/they] is/are running'). All of these moods bear no formal markings for person and number in Latvian, but can be used as verbal predicates in sentences just as any other finite verbs.

3.3. Syntactic annotation

The syntactic annotation in UD is based on the Universal Stanford Dependencies [9]. To construct the syntactic annotation, we have manually defined a relation between roles provided in UD guidelines and syntactic labeling used in Latvian Treebank. This relation is rather complicated as no labels could be aligned one to one, and morphological information (tag, sometimes also lemma or form) and local tree structures must also be taken into account. For example, according to morphological information in Latvian Treebank the role *obj* is mapped to either *obj* or *dobj* in a UD treebank, while multiple kinds of adverbial clauses (time, place, manner, etc.) are all mapped to *advcl*. Full relation between dependency roles in Latvian Treebank and UD is given in the Figure 1.

Transforming phrase style constructions requires not only assigning a dependency role to each constituent, but also creating dependency links between them. Figures 2 to 5 demonstrate how dependency labels are assigned for each phrase part. For a PMC creating dependency links between constituents (relinking) is rather simple — all punctuation marks and any other element is made dependent of the node labeled *basElem* — this is the node invoking the use of these punctuation marks. For coordination constructions relinking is done by making all other elements children of the first conjunct.

For x-words relinking is done according to x-word type: for *xApp* (appositional construction) the last element is made root; for *xPrep* (prepositional construction), *xSimile* (simile formed from conjunction and nominal) and *xParticle* (particle modified nominal) the element with role *basElem* (nominal) is made root; for compounding constructions *subrAnal*, *coordAnal, phrasElem* and named entity construction *namedEnt* the first element is made root.

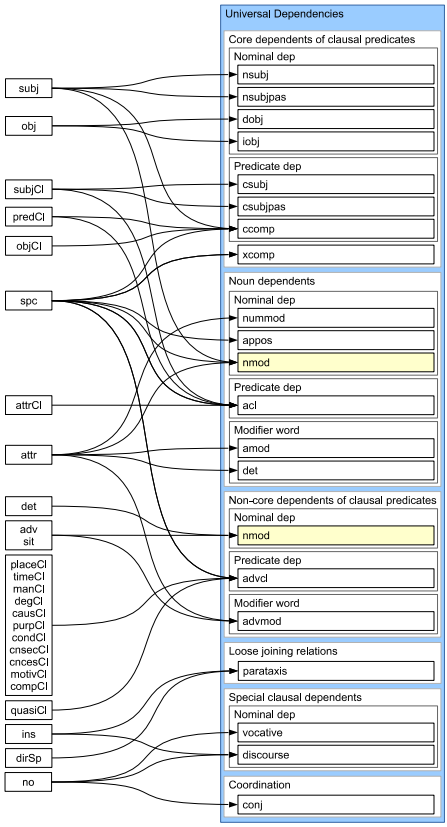

Figure 1. Transforming dependency relations.

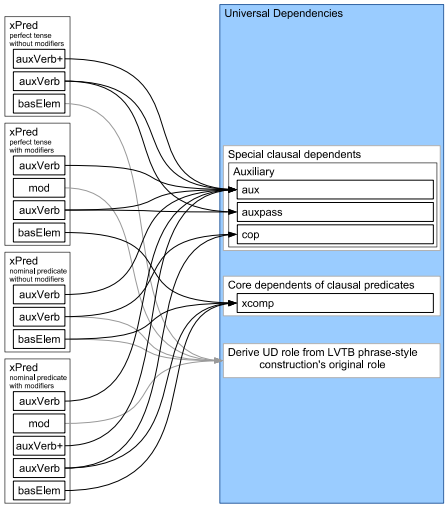

Figure 3. Assigning roles for parts of x-words — complex predicates.

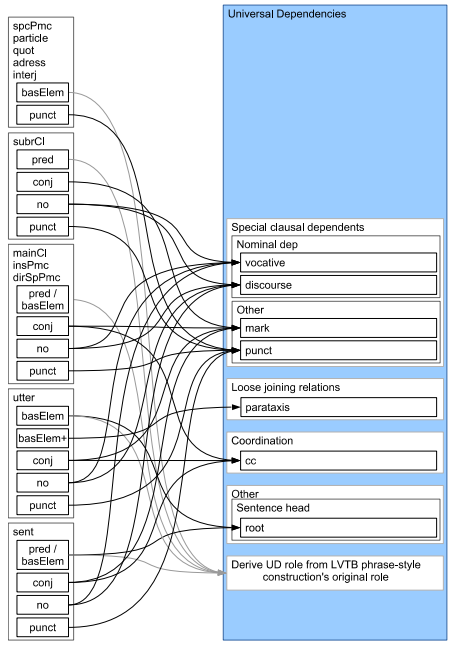

Figure 2. Assigning roles for parts of PMCs.

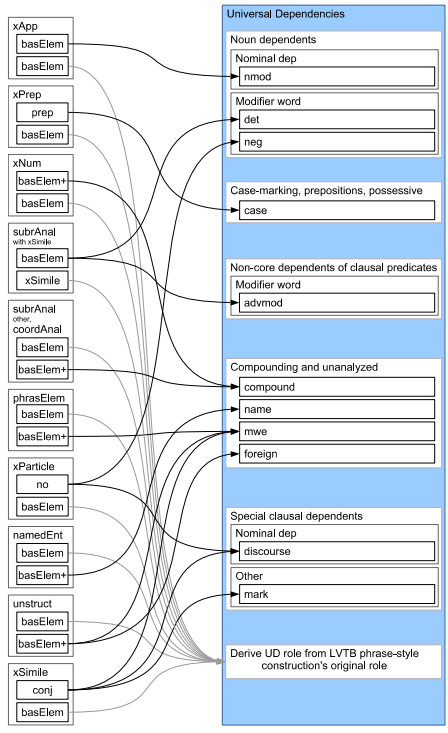

Figure 4. Assigning roles for parts of other x-words.

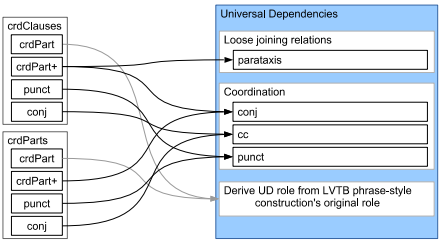

Figure 5. Assigning roles for parts of coordination.

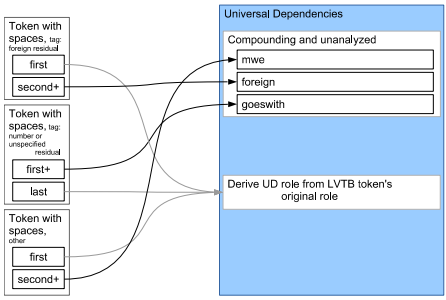

Figure 6. Assigning roles for parts of the "words with spaces".

More complicated treatment is needed for *subrAnal* in form of *vairāk* + *xSimile* (*vairāk nekā pieklājīgs* 'more than polite') or *tāds* +*xSimile* (*tāds kā nepabeigts* /such like unfinished/ 'kind of unfinished') — in these cases the *basElem* from *xSimile* is made root and *tāds/vairāk* dependent of it. As an exception, the simile conjunction *kā* 'like' / *nekā* 'unlike' is made dependent of *vairāk* 'more' in the case of *vairāk* + *xSimile* to annotate this construction similarly as in English *more than* is annotated.

The most complicated processing is needed for *xPred* phrases — complex predicates. To obtain an adequate UD representation for *xPred* the conversion procedure must take into account the following considerations:

- if *xPred* represents a nominal predicate, the copula forms the root of the phrase representing the subtree;
- if *xPred* represents a perfect tense, the main verb forms the root of the subtree;
- if *xPred* contains a modal modifier, then the verb it modifies is *xcomp* (if multiple modifiers are present, the chain of *xcomp* is formed).

Currently correct UD representation both in terms of roles (see Figure 3) and structure can be obtained only for either short *xPred*s (two constituents) or for *xPred*s with neutral word order, as the current Latvian Treebank annotation does not indicate precise relations between all parts of a complicate *xPred*.

During the development of the conversion procedure, we identified several problems that are hard or impossible to tackle using only annotations available in the current version of Latvian Treebank.

The most important among these problems is the distinction between UD roles *xcomp* and *ccomp* expressed with an infinitive verb. In Latvian Treebank in both samples *viņš pavēlēja rakt* /he order.PST3 dig.INF/ 'he ordered [someone] to dig' and *es atgriezos strādāt* /I return.PST3 work.INF/ 'I returned to work' the underlined part is labeled as *spc* (secondary predicative component) with no further distinction. Meanwhile, the UD annotation scheme requires to make a distinction based on whether it is possible for the secondary predicate to have a subject not coinciding with the subject of the main predicate, thus, making *rakt* 'to dig' in the first example *ccomp* and *strādāt* 'to work' in the second example *xcomp*. Currently in all such cases *ccomp* is assigned.

Other problems are related to ellipsis annotation. In Latvian Treebank an omitted dependency head is represented with a "fake" node, and the dependents are attached with their regular roles to this node. However, in UD, if ellipsis is done to reduce redundancy, e.g,, *Peter went to Paris, Miriam — to Prague* role *remnant* is used and

the dependent with the missing head is attached to the word fulfilling the same role in the clause where the omitted word originally occurs (*Miriam* is dependent of *Peter*, *Paris* — to *Prague*). Automatic conversion in such cases has two sources of errors: (1) in multiclause sentences it is hard to determine in which clause the omitted word first occurs, and (2) in case of multiple dependents with the same role, e.g. various adverbials, it is hard to determine the exact head for the *remnant*. Currently all such trees (3,3%) were omitted from the Latvian UD Treebank, but we plan to include them in the next version. It has to be admitted that some of the above-described problems requires changing or extending annotation in the original Latvian Treebank before we can obtain a fully UD-compatible Latvian UD Treebank.

Several relations described in UD guidelines are not used in Latvian Treebank due to various reasons:

- *reparandum* is not used, because current data contains no such text fragments. This relation will be put in use when Latvian UD Treebank will be extended with colloquial conversations or similar texts.
- *list* and *remnant* are not used, because these phenomena are not annotated in Latvian Treebank. Lists in the scope of a single sentence are annotated as coordinations.
- *dislocated* and *expl* are not used, because we do not find these roles relevant for Latvian. *dislocated* is not relevant due to rather free word order — any grammatical dislocation can be assigned a normal sentence role, e.g., *nsubj* or *dobj*. *expl* is not relevant as Latvian does not use syntactic expletives (again, due to rather free word order). However, if in any future release reflexive verbs will be split into verbs and reflexive pronouns (see Section 3.1) *expl* will be used.

4. Conclusion

The main result of this work is the first Universal Dependency treebank for Latvian. It consists of approx. 1K sentences from newswire texts. Latvian UD Treebank was published in the Universal Dependency Treebank version 1.3. It is publicly available through the LINDAT catalogue[4] and the Universal Dependencies GitHub repository[5]. The conversion code is also made publicly available[6].

After releasing the first version of Latvian UD Treebank, we made comprehensive analysis on constructions where the above described conversion process was not able to deliver a suitable UD representation. This allowed us to both correct data errors and enrich the transformation procedure, thus, enabling us to deliver better quality data for the next UD releases.

We also consider it important to work on including more data from Latvian Treebank. Currently the SyntaxNet based UD parser [5] for Latvian reports 58.92% UAS and 51.47% LAS[7], while previous studies on Latvian Treebank report up to 75.13–76.81% UAS and 65.70–66.82% LAS, depending on the annotation model used

[4] LINDAT handle http://hdl.handle.net/11234/1-1699

[5] Repository for Latvian https://github.com/UniversalDependencies/UD_Latvian

[6] Available as a part of the Latvian Treebank toolkit https://github.com/LUMII-AILab/CorporaTools; is subject to occasional improvements.

[7] Experiment results from Google Research https://github.com/tensorflow/models/blob/master/syntaxnet/universal.md

[10]. These results are not directly comparable as they use different parsers and different dependency annotation styles, however the notable difference suggests that creating more UD data should lead to some parsing accuracy improvement.

An open question regarding further Latvian treebanking is how to maintain balance between annotations familiar to the local Latvian linguists (current annotation model used in Latvian Treebank) and the aim to make our data internationally available with the help of the UD framework.

References

[1] J. Nivre, M.C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C.D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, D. Zeman, *Universal Dependencies v1: A Multilingual Treebank Collection*. Proceedings of the 10th International Conference on Language Resources and Evaluation (2016), 1659–1666.

[2] R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N.B. Castelló, J. Lee. *Universal Dependency Annotation for Multilingual Parsing.* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (2013), pp. 92–97.

[3] 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 (2011), 119–127.

[4] L. Pretkalniņa, L. Rituma, *Syntactic issues identified developing the Latvian treebank*. Proceedings of the 5th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 247, 2012, 185–192.

[5] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, M. Collins, Globally Normalized Transition-Based Neural Networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016), 2442–2452.

[6] P. Paikens, L. Rituma, L. Pretkalnina, Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics* (2013), 267–277.

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

[8] A. Kalnača, I. Lokmane, The semantics and distribution of Latvian reflexive verbs, *Multiple Perspectives in Linguistic Research on Baltic Languages* (2012), 229–256.

[9] M.C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, C. Manning, Universal Stanford Dependencies: A cross-linguistic typology, *Proceedings of 9th International Conference on Language Resources and Evaluation* (2014), 4585–4592.

[10] L. Pretkalniņa, L. Rituma, *Constructions in Latvian Treebank: the Impact of Annotation Decisions on the Dependency Parsing Performance.* Proceedings of the 6th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 268, 2014, 219–226.