Some steps towards the UNIDIVE BLARK: revising Latvian language resources and tools

Inguna Skadiņa[1,2], Baiba Saulīte[2], Laura Rituma[2], and Lauma Pretkalniņa[2]

[1]Tilde, Vienibas gatve 75a, Riga, Latvia, e-mail: inguna.skadina@tilde.lv
[2]Institute of Mathematics and Computer Science, University of Latvia, Raiņa 29, Riga, Latvia
e-mail: [firstname.lastname]@lumii.lv

Relevant UniDive working groups: WG1, WG2 and WG3

1 Introduction

The UniDive aims "to reconcile language diversity with rapid progress in language technology."[1] by investigation and application of three measures: (i) NLP-applicable universality of terminologies and methodologies, (ii) quantifying interintra-linguistic diversity, (iii) universality- and diversity-driven development of language resources and tools". In our abstract we address the measure of universality and diversity-driven development of language resources and tools (LRTs). To obtain a comprehensive understanding of the current situation, we propose to start with the construction of the Basic Language Resource Kit (BLARK) (Maegaard et al., 2006) for UniDive LRTs that are in focus of WG1-WG3, concentrating on languages represented by the consortium.

This abstract provides an overview of the Latvian language technologies and tools that could serve and be extended for the UniDive goals, in particular, corpora (WG1), lexicons (WG2) and tools (WG3). The Latvian language is a morphologically rich language with rather free word order. There are about 1.5 million native speakers of Latvian.[2] It is often called less-resourced, however, for several LRT groups that are in focus of UniDive Latvian is rather well represented.

In addition, we share our ideas on how Latvian LRTs could be linked to the corresponding resources in other languages to reach the primary goals of this COST action.

2 Corpora and corpus annotation

Latvian National Corpora Collection (LNCC) is a diverse collection of corpora representing both written and spoken language (Saulite et al., 2022a). Currently, it consists of 33 corpora, including Balanced Corpus of Modern Latvian (LVK2022; 100 million words), Latvian Treebank (LVTB; around 17k sentences) and Latvian CommonCrawl corpus (492.6M tokens). 27 corpora (total size 2.3 billion tokens) are included in the federated search, most of these corpora (23) are automatically annotated by the open-source morphological tagger (Paikens, 2016). The common tagset of Latvian generally complies with the MULTEXT-EAST standard (Erjavec, 2017), adapted to the Latvian specifics.

Latvian Treebank is manually annotated using a hybrid dependency-constituency grammar model (Barzdins et al., 2007). LVTB is also transformed to the Universal Dependencies model (Latvian UD Treebank) (Pretkalnina et al., 2018) and is available for cross-lingual research.

The Latvian multilayer corpus FullStack-LV (Gruzitis et al., 2018b) is anchored in the following cross-lingual state-of-the-art representations: Universal Dependencies, FrameNet, PropBank and Abstract Meaning Representation (AMR). A part of FullStack-LV, Latvian FrameNet corpus (Gruzitis et al., 2018a) is annotated using the latest inventory of Berkeley FrameNet. Currently, 570 Berkeley FrameNet frames have been used for semantic annotation of the Latvian FrameNet corpus, 2900 lexical units and almost 26 000 usage examples have been annotated (Saulite et al., 2022a).

Unlike the FrameNet project, in which semantic annotation was done for a specific instance in a sentence (target word) and its associated semantic roles, manual semantic annotation is currently underway, in which all words in a sentence are semantically annotated. The goal is to create a gold standard corpus for the evaluation of the word sense disambiguation (WSD) solutions.

3 Lexicons and corpus interfaces

On-line dictionary for Latvian *Tezaurs.lv* (Spektors et al., 2016) is a large lexical dataset (around 388k entries). It has been extended with structured data for various NLP needs (inflectional paradigm

[1]Memorandum of Understanding (MoU): https://e-services.cost.eu/files/domain_files/CA/Action_CA21167/mou/CA21167-e.pdf
[2]https://valoda.lv/

and inflection tables) and gives lemmas and other grammatical features for the morphological tagger. The data is prepared in TEI format.[3]

Tezaurs.lv contains a large number of MWEs, however, the structural analysis and annotation of MWEs has only just started. It focuses on the syntactic analysis and syntactic patterns of MWEs. It is planned to consider the lexical and frame semantic aspects and patterns of a selected subset of MWEs as well.

Another lexical resource, implemented on the basis of the *Tēzaurs.lv* platform, is Latvian WordNet (Paikens et al., 2022), in which synsets and semantic links are created for the most frequently used words. These synsets are aligned with Princeton WordNet synsets opening the possibility to include Latvian WordNet in multilingual resources, for example, Open Multilingual Wordnet. The data is also available in LMF format.[4]

For the needs of Latvian Wordnet, the *Tezaurs.lv* editor's tool has been adapted for the selection and linking of corpus examples to a specific word and MWE sense. As result, a large amount of corpus examples has been collected in the thesaurus database (about 70,000), which will be used in WSD experiments in the future.

Latvian terminology is consolidated in the Latvian National Terminology Portal[5] and integrated into the European Terminology Bank EuroTermBank (Vasiljevs et al., 2008). EuroTermBank contains about 3.5 million entries from 463 collections in 44 languages. This database uses unified data exchange mechanisms based on the latest versions of the TBX standard.[6]

4 Language processing tools

The UniDive efforts are concentrated around three groups of language processing tools: syntactic parsers, semantic parsers and MWE processing tools for discovery and identification.

Latvian **syntactic parsers** have been trained mainly on UD treebanks (Znotins and Barzdins, 2020) and thus could be easily integrated into universal solutions.

Where it concerns **semantic parsers**, several experiments have been performed with the Grammatical Framework (Gruzitis et al., 2012), while Latvian FrameNet is used to generate Latvian (Gruzitis et al., 2020; Saulite et al., 2022b). Moreover, several experiments address generation to and from AMR, thus providing means and supporting interlingual and cross-lingual semantic parsing (Znotiņš et al., 2020).

While syntactic and semantic parsers for Latvian already today mostly follow common standards, tools for **MWE identification** are not so well developed and thus do not always follow common standards. Several experiments were performed during PARSEME COST action and continued through the FullStack-LV project (Skadina, 2018).

Most recent work is related to named entity recognition (NER, Znotins and Barzdins 2020, Vīksna and Skadiņa 2020) and terminology identification and cross-lingual alignment. For NER, good results have been demonstrated not only in monolingual settings but also for cross-lingual NE linking between Slavic languages (Vīksna and Skadina, 2021). Both, named entities and terminology items are annotated with BIO markup (Ramshaw and Marcus, 1995).

5 Conclusion

In this abstract we provided an overview of Latvian language resources and tools and discussed their compliance with the universality goal of the UniDive action.

We can conclude that Latvian resources and tools are mainly developed in accordance with international standards and models (e.g. TEI, MULTEXT-EAST, LMF, TBX, UD, FrameNet, AMR, etc.), thus support cross-lingual studies of universality and diversity. While MWE resources are less developed today, some recently started initiatives, e.g. LATE and DHELI projects, could allow us to fill this gap soon.

Acknowledgments

The results reported here were supported by COST action "Universality, diversity and idiosyncrasy in language technology" (UNIDIVE, CA21167), the project "Research on Modern Latvian Language and Development of Language Technology" (LATE, project No: VPP-LETONIKA-2021/1-0006) and project "Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia" (DHELI, project No: VPP-IZM-DH-2022/1-0002).

[3] https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

[4] https://globalwordnet.github.io/schemas/#xml

[5] https://termini.gov.lv/

[6] https://www.tbxinfo.net/

References

G. Barzdins, N. Gruzitis, G. Nespore, and B. Saulite. 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)*.

T. Erjavec. 2017. *MULTEXT-East*, pages 441–462. Springer Netherlands, Dordrecht.

N. Gruzitis, R. Dargis, L. Rituma, G. Nespore-Berzkalne, and B. Saulite. 2020. Deriving a PropBank Corpus from Parallel FrameNet and UD Corpora. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 63–69.

N. Gruzitis, G. Nespore-Berzkalne, and B. Saulite. 2018a. Creation of Latvian FrameNet based on Universal Dependencies. In *Proceedings of the International FrameNet Workshop (IFNW)*, pages 23–27.

N. Gruzitis, P. Paikens, and G. Barzdins. 2012. FrameNet resource grammar library for GF. In *Controlled Natural Language*, volume 7427. Springer.

N. Gruzitis, L. Pretkalnina, B. Saulite, L. Rituma, G. Nespore-Berzkalne, A. Znotins, and P. Paikens. 2018b. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 4506–4513.

B. Maegaard, S. Krauwer, K. Choukri, and L. Jørgensen. 2006. The BLARK concept and BLARK for Arabic. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

P. Paikens. 2016. Deep neural learning approaches for Latvian morphological tagging. In *Human Language Technologies - The Baltic Perspective*, volume 289. IOS Press.

P. Paikens, M. Grasmanis, A. Klints, I. Lokmane, L. Pretkalnina, L. Rituma, M. Stade, and L. Strankale. 2022. Towards Latvian WordNet. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 2808–2815.

L. Pretkalnina, L. Rituma, and B. Saulite. 2018. Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank. In *Text, Speech, and Dialogue*, volume 11107, pages 95–105. Springer.

L. Ramshaw and M. Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*.

B. Saulite, R. Dargis, N. Gruzitis, I. Auzina, K. Levane-Petrova, L. Pretkalnina, L. Rituma, P. Paikens, A. Znotins, L. Strankale, K. Pokratniece, I. Poikans, G. Barzdins, I. Skadina, A. Baklane, V. Saulespurens, and J. Ziedins. 2022a. Latvian National Corpora Collection – Korpuss.lv. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 5123–5129.

B. Saulite, G. Nespore-Berzkalne, L. Rituma, V. Lasmanis, and N. Gruzitis. 2022b. Latviešu valodas FrameNet korpuss. *Letonica*, (47):284–296.

I. Skadina. 2018. Looking for a Needle in a Haystack: Semi-automatic Creation of a Latvian Multi-word Dictionary from Small Monolingual Corpora. In *Proceedings of the 18th EURALEX International Congress*, pages 255–265.

A. Spektors, I. Auzina, R. Dargis, N. Gruzitis, P. Paikens, L. Pretkalnina, L. Rituma, and B. Saulite. 2016. Tezaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

A. Vasiljevs, S. Rirdance, and A. Liedskalnins. 2008. EuroTermBank: Towards greater interoperability of dispersed multilingual terminology data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources ICGL*, pages 213–220.

R. Vīksna and I. Skadina. 2021. Multilingual Slavic Named Entity Recognition. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 93–97, Kiyv, Ukraine. Association for Computational Linguistics.

R. Vīksna and I. Skadiņa. 2020. Large Language Models for Latvian Named Entity Recognition. In *Baltic HLT*.

A. Znotins and G. Barzdins. 2020. LVBERT: Transformer-based model for Latvian language understanding. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press.

A. Znotiņš, P. Paikens, and N. Grūzītis. 2020. Latvian AMR sembank. CLARIN-LV digital library at IMCS, University of Latvia.