The Contribution of PARSEME to enhancing the Electronic Dictionary *Tēzaurs*

Gunta Nešpore-Bērzkalne
Institute of Mathematics and Computer Science, University of Latvia
Raiņa blvd. 29, Riga, Latvia
gunta.nespore@lumii.lv

Relevant UniDive working groups: WG1, WG2

1 Introduction

Latvian is one of the languages included in the PARSEME 2.0 multilingual corpus annotated for multi-word expressions (MWEs). The corpus covers MWEs of all syntactic categories and has been annotated according to cross-lingually unified guidelines. The Latvian corpus contains texts of different genres from the UD LVTB treebank. There are 3,203 annotated MWEs in 13,784 sentences (Savary et al., 2026).

For Latvian, this is the first publicly available manually annotated MWE corpus. Up until now there have been relatively few studies of Latvian MWEs, some of which have been carried out to accommodate machine translation (see e.g. Deksne et al. (2008); Skadiņa (2018)). In Latvian linguistics, mainly idiomatic MWEs (phraseology) have been studied and described, with the main focus on their meaning and figurativeness, not on the formal structure or syntactic types (see e. g. Laua (1992); Veisbergs (2019); Rituma et al. (2024)).

The main MWE resources available for Latvian so far are dictionaries (mainly intended for human use), with the electronic dictionary *Tēzaurs* as the most exhaustive such resource.

The purpose of the abstract is to outline how experience and data gained from annotating Latvian PARSEME corpus is planned to be implemented for the enhancement of *Tēzaurs*.

2 Electronic dictionary *Tēzaurs*

Tēzaurs[1] is the largest Latvian monolingual electronic dictionary comprising more than 410,000 entries. It was started as a compilation from nearly 300 pre-existing dictionaries and other sources and has since developed into a multifunctional lexical resource consisting of granular interlinked entities enriched with various kinds of metadata. Besides entries for single words, *Tēzaurs* also contains ca. 76,000 entries for MWEs (Grasmanis et al., 2023), making it currently the most comprehensive resource of MWEs in Latvian lexicography.

MWE entries of *Tēzaurs* may contain the same elements as single-word entries, such as lexical variants of MWE, tags of style and usage context, sense definitions, a section for adding examples from corpora. Additionaly, MWE entries are assigned MWE category tags (e.g., phraseological unit, collocation, multi-word proper noun, complex term) (Rituma et al., 2024). However, MWE entries in *Tēzaurs* lack explicit information on morphological and syntactic MWE structure.

As structurally independent units, MWE entries can be linked with other entries in *Tēzaurs*, including entries of their component-words (this is how MWEs are displayed in single-word entries: the entry of the respective word shows all MWEs that the word is linked with), as well as entries involved in the Latvian WordNet (Paikens et al., 2023) showing semantic relations between MWEs and other lexical units. They can also be linked with the Princeton English WordNet (Fellbaum, 1998).

3 Enhancing *Tēzaurs*

During work on the PARSEME corpus, Annotation guidelines (version 2.0)[2] were applied to Latvian data. As expected, the annotation of certain syntactic types of MWEs did not pose significant difficulties; however, the applicability of some categories to Latvian requires further investigation. Nevertheless, the resulting data constitute a valuable resource, and it is planned to use them to enhance *Tēzaurs*. This includes both a quantitative expansion of the dictionary (Subsections 3.1 and 3.2) and a change of approach to certain types of word combinations (Subsection 3.3).

3.1 Adding new MWEs

The MWEs annotated in the PARSEME corpus meet the MWE criteria set out in the Annotation

[1] Publicly available interactively at https://tezaurs.lv or as data from https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/156

[2] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0//

guidelines (they show idiosyncrasy in some aspect), thus, their inclusion in the explanatory dictionary should also be considered. Fortunately, we find that the vast majority of the annotated MWEs are already represented in *Tēzaurs*. However, some are still missing, for instance, more recently coined MWEs, e. g., *zaļais tirgus* (lit. 'the green market') 'farmers market'. A notable group of constructions not represented in *Tēzaurs* is light verb constructions (LVCs). In Latvian linguistics light verbs are not acknowledged as a distinct class since LVCs have traditionally been considered uncharacteristic for Latvian. Consequently, they have neither been studied in detail nor systematically included in dictionaries. Some of them have been discussed from the perspective of language standardization, where their usage is regarded as undesirable (Freimane, 1993, 113-118). However, the PARSEME corpus data demonstrate, that LVCs are more common in Latvian than previously assumed (277 annotated instances of LVC in the corpus)[3], indicating that these constructions would merit from further investigation and that at least some of them should be included in the dictionary.

A list of MWEs in Latvian PARSEME corpus will be extracted and compared with the *Tēzaurs* data to identify MWEs that are not yet represented and new MWE entries will be created. Furthermore, the corpus data will be examined to determine whether respective MWEs are used in senses that are defined in *Tēzaurs*; where necessary, the corresponding entries will be supplemented with new senses.

3.2 Adding PARSEME corpus sentences to MWE entries as corpus examples

In the *Tēzaurs* platform, corpus examples for dictionary entries are primarily added from the Latvian National Corpora Collection, which comprises 45 corpora developed by 15 institutions[4] (Saulīte et al., 2022), including the UD LVTB treebank, on which the PARSEME corpus is based. It is planned to add the PARSEME corpus sentences containing MWEs to the corresponding MWE entries. This may be achieved either by storing copies of the sentences within *Tēzaurs* or by creating link to an external resource. Further technical research is needed to determine the best approach.

[3] https://gitlab.com/parseme/corpora/-/wikis/home/

[4] https://korpuss.lv/en/about/

3.3 Redefining certain expressions as MWEs

A part of the expressions that match the PARSEME MWE criteria (e. g., contain cranbery word, irregular syntactic structure, prohibited modification[5]) have not been described in Latvian linguistics as MWEs. These are mostly functional expressions that are not particularly idiomatic, often with a very general lexical meaning. They often correspond to functional MWEs – conjunction idioms and interjection idioms – but in some cases – MWEs of other types as well. E. g., *kaut gan* 'although'; *kā arī* 'as well as'; *aijā, žūžū* (lullaby refrain) 'rock-a-bye'; *kling, klang* (lit. 'cling clang') 'dong-dong'; *čumēt un mudžēt* (lit. 'to surge and swarm') 'describes a situation when a large mass of people or animals is moving'.

In Latvian linguistics, such expressions are described mainly from a grammatical point of view – as a conjunctions, particles or interjections, citing "that conjunctions can consist of several words" or "particles often occur in combinations" (Kalnača and Lokmane, 2021, 362, 375).

Thus, in Latvian explanatory dictionaries such word combinations are typically not described as separate units, but rather as one of the senses in one or more entries of their component words. For instance, *kaut gan* 'although' is described as one of the senses of the conjunction *kaut* 'though', which functions in combination with the particle *gan* 'emphasis, e.g., really'. Or *aijā, žūžū* 'rock-a-bye' is described as the only meaning of the interjection *žūžū* (a cranberry word), which functions in the corresponding word combination. This approach is not efficient in a modern dictionary, as such expressions can neither be searched easily, nor processed automatically.

Encouraged by the experience in PARSEME, we decided to regard all word combinations described in this way as MWEs, and convert them to MWE entries, thereby integrating them into *Tēzaurs* as independent units. A systematic review of the *Tēzaurs* data from this perspective has revealed that the number of word combinations incorporated in sense definitions is substantial, amounting to more than 800 instances. Their restructuring represent a significant improvement of the resource. The first new MWE entries have already been created, e.g., for the word combination (conjuntion idiom

[5] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0//

in PARSEME) *kā arī* 'as well as' [6].

4 Conclusion

The experience gained during the development of the PARSEME corpus was valuable and beneficial, as it provided an opportunity to view Latvian language data from a different perspective. This experience, along with the corpus data, will be utilized both for the improvement of *Tēzaurs* and for future research on Latvian multi-word expressions.

Acknowledgements

The research was supported by the State Research program Letonika – Fostering a Latvian and European Society; project Digital Resources and AI Technologies for the Sustainability of the Latvian Language (VPP-IZM-LETONIKA-2025/1-0004)

References

Daiga Deksne, Raivis Skadiņš, and Inguna Skadiņa. 2008. Dictionary of multiword expressions for translation into highly inflected languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. Mit Press.

Inta Freimane. 1993. *Valodas kultūra teorētiskā skatījumā*. Zvaigzne, Rīga.

Mikus Grasmanis, Pēteris Paikens, Lauma Pretkalniņa, Laura Rituma, Lauma Strankale, Artūrs Znotiņš, and Normunds Grūzītis. 2023. Tēzaurs.lv – the experience of building a multifunctional lexical resource. In *Electronic lexicography in the 21st century (eLex): Invisible Lexicography*, pages 400–418.

Andra Kalnača and Ilze Lokmane. 2021. *Latvian Grammar*. University of Latvia Press, Riga.

Alīse Laua. 1992. *Latviešu valodas frazeoloģija*. Zvaigzne, Rīga.

P. Paikens, A. Klints, I. Lokmane, L. Pretkalnina, L. Rituma, M. Stade, and L. Strankale. 2023. Latvian wordnet. In *12th Global Wordnet Conference*, pages 187–196. Global Wordnet Association.

Laura Rituma, Gunta Nešpore-Bērzkalne, Agute Klints, Ilze Lokmane, Madara Stāde, and Pēteris Paikens. 2024. Classifying multi-word expressions in the latvian monolingual electronic dictionary tēzaurs.lv. In *6th International Conference Computational Linguistics in Bulgaria (CLIB)*, pages 113–118.

Baiba Saulīte, Roberts Darģis, Normunds Grūzītis, Ilze Auziņa, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Pēteris Paikens, Artūrs Znotiņš, Lauma Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Bārzdiņš, Inguna Skadiņa, Anda Baklāne, Valdis Saulespurēns, and Jānis Ziediņš. 2022. Latvian national corpora collection – korpuss.lv. In *13th Language Resources and Evaluation Conference (LREC)*, pages 5123–5129.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiș, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, Chaya Liebeskind, Irina Lobzhanidze, Aleksandra Marković, Gunta Nešpore-Bērzkalne, Adriana Pagano, Mehrnoush Shamsfard, Ranka Stanković, Vahide Tajalli, Carole Tiberius, and Aakanksha Padhye. 2026. Parseme 2.0 multilingual corpus of multiword expressions. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834.

Inguna Skadiņa. 2018. Looking for a needle in a haystack: Semi-automatic creation of a latvian multiword dictionary from small monolingual corpora. In *Proceedings of the 18th EURALEX International Congress*, pages 255–265.

Andrejs Veisbergs. 2019. The fuzzy concept of idiom and what it might mean for bilingual dictionaries. *Baltic Journal of English Language, Literature and Culture*, 9:111–129.

[6] https://tezaurs.lv/mwe:393437/