Project proposition for the 1st UniDive Training School, 8–12 July 2024

Integrating PARSEME-MWEs into the Spoken Slovene SST UD Corpus

Applicant: Tina Munda, Centre for Language Resources and Technologies at the University of Ljubljana (CJVT UL), Slovenia

Keywords: multi-word expressions, spoken, speech

Slovene has participated in the PARSEME shared task since its beginning in 2017 (Savary et al., 2017) and saw the pilot application of the PARSEME-MWE scheme in the training corpus ssj500k v2.0 (Krek et al., 2017), quickly followed by an upgrade according to the enriched guidelines 1.1[1] in the next version of the ssj500k corpus – v2.1 (Krek et al., 2018). Although the Slovene team has remained an active participant in the initiative, now part of the UniDive shared task, no new Slovene data has been annotated since.

For the 1st UniDive Training School, I propose applying the PARSEME-MWE annotation scheme to a corpus of spoken Slovene – SST (Spoken Slovenian Treebank) UD (Dobrovoljc & Nivre, 2016). The original version of SST UD is a sample of the reference corpus of spoken Slovene – Gos (Zwitter Vitez et al., 2013), which comprises "audio recordings and transcripts of approximately 120 hours (1 million words) of monologic, dialogic and multi-party spontaneous speech in different everyday situations" (Dobrovoljc & Nivre, 2016). The corpus has recently been significantly upgraded. According to the project README, it "has been partially revised and substantially extended with new data from GOS v2 (Verdonik et al., 2024), such as parliamentary debates, round tables and online events. The latest version of the SST treebank thus includes 6,104 utterances (76,341 tokens), produced by 676 speakers in 344 different speech events (48% public and 52% non-public tokens)" (Dobrovoljc, 2024).

My participation in the training school would initiate work on a new corpus, which would not only provide a new MWE resource for Slovene but also represent the first *speech* corpus of Slovene containing PARSEME-MWEs, and to my knowledge, the first *speech* corpus ever annotated with the PARSEME-MWE scheme.

In light of this, the questions for the planned brainstorming hackathon that arise are:
What are the specific challenges of annotating MWEs in speech corpora compared to written ones?
Do MWEs identified in speech differ from those in written language?
Are there tools for automatic MWE annotation? If so, are they language-specific or universal? Are there estimations on how many manual annotations are needed to achieve a highly accurate machine pre-annotation of MWEs?

I am excited about the prospect of participating in the training school, especially after the UniDive webinar for newcomers last June, which sparked my interest in MWEs. While I am already familiar with Universal Dependencies for Slovene, including part-of-speech and syntactic annotation, I am eager to learn more about the PARSEME-MWE guidelines, the corpus-annotation infrastructure, and the community. To sum up, the goal of my participation would be to start annotating MWEs in the proposed corpus, with the long-term goal of publishing the enhanced SST corpus as a resource for the wider research community.

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/

References:

Dobrovoljc, K., & Nivre, J. 2016. *The Universal Dependencies Treebank of Spoken Slovenian*. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia.

Dobrovoljc, K. (2024). README.md for UniversalDependencies/UD_Slovenian-SST. GitHub. https://github.com/UniversalDependencies/UD_Slovenian-SST/tree/dev?tab=readme-ov-file

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., & Kuzman, T. (2017). *Training corpus ssj500k 2.0*. http://hdl.handle.net/11356/1165

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., & Zajc, A. (2018). *Training corpus ssj500k 2.1*. http://hdl.handle.net/11356/1181

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., & Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In S. Markantonatou, C. Ramisch, A. Savary, & V. Vincze (Eds.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 31–47). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1704

Verdonik, D. et al. (2024). Gos 2: A New Reference Corpus of Spoken Slovenian. *LREC-Coling 2024 Proceedings*.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2013). *Spoken corpus Gos 1.0*. http://hdl.handle.net/11356/1040