# Extension of the ELEXIS-WSD Parallel Sense-Annotated Corpus Within UniDive: New Languages and Layers

**Jaka Čibej**, University of Ljubljana, Slovenia, jaka.cibej@ff.uni-lj.si
**Simon Krek**, University of Ljubljana, Slovenia, simon.krek@ff.uni-lj.si
**Carole Tiberius**, Dutch Language Institute, The Netherlands, carole.tiberius@ivdnt.org
**Irina Lobzhanidze**, Ilia State University, Georgia, irina_lobzhanidze@iliauni.edu.ge
**Verginica Barbu Mititelu**, Research Institute for Artificial Intelligence, Romania, vergi@racai.ro
**Jelena Kallas**, Institute of the Estonian Language, Estonia, jelena.kallas@eki.ee
**Kertu Saul**, Institute of the Estonian Language, Estonia, kertu.saul@eki.ee
**Kadri Muischnek**, University of Tartu, Estonia, kadri.muischnek@ut.ee
**Karoliina Jõgi**, University of Tartu, Estonia, karoliina.jogi@ut.ee
**Ranka Stanković**, University of Belgrade, Serbia, ranka@rgf.rs
**Cvetana Krstev**, Association for Language Resources and Technologies, Serbia, cvetana@jerteh.rs
**Aleksandra Marković**, Institute for the Serbian Language SASA, Serbia, aleksandra.markovic@isj.sanu.ac.rs
**Ana Ostroški Anić**, Institute of Croatian Language and Linguistics, Croatia, aostrosk@ihjj.hr
**Bartłomiej Alberski**, Wrocław University of Science and Technology, Poland, bartlomiej.alberski@pwr.edu.pl
**Vladimir Cvetkoski**, Ss. Cyril and Methodius University, North Macedonia, cvetkoski@flf.ukim.edu.mk
**Voula Giouli**, Aristotle University of Thessaloniki, Greece, pgiouli@del.auth.gr
**Rusudan Makhachashvili**, Borys Grinchenko Kyiv Metropolitan University, Ukraine, r.makhachashvili@kubg.edu.ua
**Olha Kanishcheva**, Heidelberg University, Germany, kanichshevaolga@gmail.com

***Relevant UniDive working groups:*** WG2, WG1

## 1 Introduction

Within UniDive (Savary et al., 2024), Task 2.2 (*Design of a lexicon-corpus interface*) of WG2 involves the development and upgrade of the *ELEXIS-WSD Parallel Sense-Annotated Corpus* (*ELEXIS-WSD* for short), a small-scale parallel corpus consisting of language subcorpora containing translations of the same sentences in multiple languages; with manual tokenization, lemmatization, morphosyntactic tagging, and, most importantly, manually assigned sense annotations from a lexicographic or lexical resource (such as monolingual dictionaries or wordnets), directly linking the corpus to the lexicon. *ELEXIS-WSD* is a useful resource for word-sense disambiguation tasks and cross-lingual comparisons (see Martelli et al., 2021 for more details of its inception and development).

The latest version before the introduction of UniDive was 1.1 (Martelli et al., 2023), with subcorpora for 10 languages: Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish.

In this paper, we present the results of the activities of Task 2.2 of *UniDive*, which culminated in the publication of version 2.0 of *ELEXIS-WSD* in April 2026.[1] The main goals included (1) the introduction of new languages to the corpus; (2) the addition of new annotation layers.

## 2 Extension Process

A total of 11 language teams participated in Task 2.2, with 8 new languages added (Romanian, Greek, Ukrainian, Polish, Macedonian, Georgian, Croatian, and Serbian), and 3 language teams working on additional annotation layers for existing subcorpora (Slovene, Estonian, Dutch). Not all language teams started the extension process at the same time and some had more human resources, so they have completed different stages of the annotation process in version 2.0. We present the extension process layer-by-layer in the following subsections.

### 2.1 Translation and Tokenization

For newly added languages, the 2,024 sentences from the English subcorpus (the original) were first translated automatically using either *Google Translate*, *DeepL*, or *GPT-4o* (depending on preference and availability). The translations were then manually corrected through cross-comparison with the original, with particular attention to mistranslations of polysemous/ambiguous words (such as *draw*) and issues that could affect other layers

[1] ELEXIS-WSD 2.0 at CLARIN.SI: http://hdl.handle.net/11356/2101

downstream (such as tokenization). The translation corrections were made in Google Sheets.

The corrected translations were then tokenized automatically using *UDPipe* (Straka et al., 2016; Straka, 2018).[2] The tokenizations were corrected manually in Google Sheets using a modified CoNLL-U format, with a particular emphasis on subtokenizations/multiwords, which were done in line with the final stage of annotation (word-sense assignment). Manual subtokenizaton corrections ensure that individual tokens or subtokens can be assigned specific senses from the sense inventory at a later stage (e.g. the Polish *chciałby* 'would like to' was subtokenized into *chciał* 'like' and *by* 'would'). Automatic consistency checks were performed at the end to ensure that the sums of all tokens correspond to entire sentences.

### 2.2 Lemmatization, Morphosyntactic Annotation and Syntactic Parsing

The manually tokenized subcorpora were then first automatically lemmatized, morphosyntactically tagged and parsed using *UDPipe*.[3] The resulting CoNLL-U files were converted to TSV3 format compatible with *INCEpTION* (Klie et al., 2018; version 23.1). Each language team received user accounts to work on the corpus, with the language leader coordinating the work if more annotators were involved. Syntactic annotations according to the Universal Dependencies system (de Marneffe et al., 2021) were not present in *ELEXIS-WSD* before UniDive. All subcorpora were parsed automatically, but due to time and resource constraints, manual corrections were optional within UniDive. Following previously outlined plans (Tiberius et al., 2024), manual corrections in version 2.0 were made for Slovene and Estonian, whereas Dutch will be completed by the end of UniDive.

### 2.3 Multiword Expressions and Named Entities

Another additional annotation layer covered multiword expressions (MWEs) according to the *PARSEME 2.0* guidelines,[4] which cover a total of 32 categories of MWEs. The guidelines were developed and used to annotate MWEs in other corpora within UniDive WG1. They were completed in mid-2025, so this activity started later. So far, only Slovene and Serbian subcorpora include MWEs in version 2.0. Estonian, Dutch, Georgian, and Romanian will also be completed by the end of UniDive.

In addition, named entities were annotated in the English subcorpus in the pre-UniDive version, but the annotations were limited to spans (without categories). Within UniDive, the Serbian language team performed the systematic annotation of named entities using a typology consisting of 8 categories (Krstev et al., 2025), identifying more than 2,350 named entity annotations and linking them to the corresponding Wikidata objects (if available). In the future, the annotations will be propagated semi-automatically to other languages to ensure that all named entities are harmonized across the corpus.

### 2.4 Sense Annotation

Among the newly added subcorpora, only Serbian has reached the stage of sense annotation. The process involves an innovative semi-automatic approach (described in more detail by Stanković et al., 2026). As this requires the preparation of a sense inventory,[5] which has turned out to be too time-consuming for most language teams involved in WG2, other subcorpora will be annotated in future endeavors (more on this in Section 4).

## 3 Version 2.0

Figure 1 shows the current state of the newly added subcorpora (as of version 2.0 of ELEXIS-WSD, published on 1st April 2026). All in all, the corpus has been extended with 8 language subcorpora and covers approx. 620,000 tokens and 38,000 sentences (an 80% increase compared to cca. 345,000 tokens and 20,000 sentences before UniDive).

## 4 Conclusion, Lessons Learned and Future Work

We have presented a new version of the *ELEXIS-WSD Parallel Sense-Annnotated Corpus* extended

[2]The following models were used: *romanian-rrt-ud-2.12-230717*, *polish-pdb-ud-2.15-241121*, *georgian-glc-ud-2.15-241121*, *croatian-set-ud-2.15-241121*, *greek-gdt-ud-2.17-251125*, *serbian-set-ud-2.15-241121*; Macedonian was annotated using Gemini 3.1 Pro via native grammar-based prompting. Ukrainian has not yet been tokenized in version 2.0.

[3]Macedonian was annotated using Gemini 3.1 Pro via native grammar-based prompting.

[4]PARSEME 2.0 Guidelines: https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

[5]A sense inventory is a lexical resource (e.g., dictionary, WordNet) containing lexemes/lexical units, their sense divisions, and their corresponding definitions.

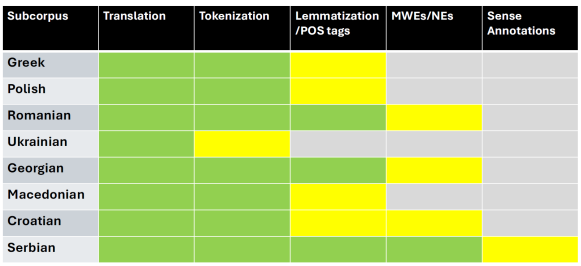


Figure 1: Current State of New Language Subcorpora Added to ELEXIS-WSD within UniDive. The green fields indicate annotation layers for which manual corrections have been completed and included in version 2.0. The yellow fields indicate layers where work is still in progress and is expected to be completed by the end of UniDive. This will be published in future versions of *ELEXIS-WSD*.

within UniDive WG2. The corpus is available under an open-access license (CC BY-SA 4.0) at the CLARIN.SI repository.[6]

Work on the corpus will continue until the end of UniDive. This will be a good starting point for the continuation of the development of *ELEXIS-WSD* within the upcoming *ELEXAI* project (*European Lexicographic Infrastructure for Artificial Intelligence*; 2026-2029). The remaining annotation layers will be completed and harmonized between subcorpora, with particular emphasis on sense annotation and sense inventories, which turned out to be a significant gap for most participating languages to overcome, as they either had no available machine-readable open-source sense inventory or only had resources that are either outdated or non-comprehensive. The next step would be to attempt to leverage large language models to generate missing senses automatically, and potentially perform more complex sense linking and supersense annotation.

## Acknowledgements

The presented work was supported by the COST Action CA21167 – *Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive), the *Estonian Research Council* grant (PRG 1978), the Polish Minister of Science (*Support for the participation of Polish scientific teams in international research infrastructure projects*, 2024/WK/01), the Science Fund of the Republic of Serbia (#7276, *Text Embeddings - Serbian Language Applications - TESLA*), the Ministry of Science of the Republic of Serbia, (451-03-33/2026-03/200174), and the Slovenian Research and Innovation Agency (research program *Language Resources and Technologies for Slovene*, P6-0411).

[6]ELEXIS-WSD 2.0 at CLARIN.SI: http://hdl.handle.net/11356/2101

## References

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Cvetana Krstev, Ranka Stanković, Aleksandra Marković, and Milica Ikonić Nešić. 2025. Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository. In *3rd General Meeting Hungarian Research Centre for Linguistics, Budapest, Hungary*.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña Ruiz, José Luis Sancho Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, and Tina Munda. 2021. Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, pages 377–395.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, and Sia Kolkovska. 2023. Parallel sense-annotated corpus ELEXIS-WSD 1.1. Slovenian language resource repository CLARIN.SI.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markan-

tonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Ranka Stanković, Cvetana Krstev, Saša Petalinkar, Milica Ikonić Nešić, Aleksandra Marković, Marina Bagi, Marijana Đukić, and Jelena Bogdanović. 2026. SrELEXIS-WSD: Hybrid Semi-Automated WSD for Serbian with Large Language Models—Results and Challenges. In *4th UniDive General Meeting*.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Carole Tiberius, Jaka Čibej, Jelena Kallas, Kertu Saul, Kadri Muischnek, and Simon Krek Krek. 2024. UD Syntax for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Pilot Study. In *UniDive 2nd General Meeting (Naples, Italy)*, Naples, Italy.