The ELEXIS-WSD Parallel Sense-Annotated Corpus and South Slavic Languages: Subcorpora for Croatian, Serbian, and Slovene

Scientific paper DOI: 10.18485/judig.2025.1.ch3

Jaka Čibej[1], 0000-0002-3037-6848
Ranka Stanković[2], 0000-0001-5123-6273
Ana Ostroški Anić[3], 0000-0001-9999-0750
Simon Krek[4], 0000-0001-8965-6863
Carole Tiberius[5] 0000-0002-7860-5427

Abstract

The open-source ELEXIS-WSD Parallel Sense-Annotated Corpus was developed within the ELEXIS project and in version 1.1 contains 2,024 sentences for each of 10 languages: Bulgarian, Danish, English, Spanish, Estonian, Hungarian, Italian, Dutch, Portuguese, and Slovene. Within the sentences, each content word (noun, adjective, verb, and adverb) has been assigned a corresponding sense from one of the 10 open-access sense inventories containing definitions. Within the context of the UniDive COST Action (CA21167), the corpus is being extended with several new languages, including South Slavic languages. In the paper, we focus on three subcorpora: Croatian, Serbian, and Slovene. We briefly describe the structure and purpose of the ELEXIS-WSD, then continue by describing

[1] Centre for Language Resources and Technologies, University of Ljubljana, Faculty of Arts, University of Ljubljana, jaka.cibej@ff.uni-lj.si
[2] University of Belgrade, Faculty of Mining and Geology, Chair for Mathematics and Informatics, ranka@rgf.rs
[3] Department of General Linguistics, Institute for the Croatian Language, aostrosk@ihjj.hr
[4] Artificial Intelligence Laboratory, "Jožef Stefan" Institute, Centre for Language Resources and Technologies, University of Ljubljana, simon.krek@ijs.si
[5] Dutch Language Institute, Centre for Linguistics, carole.tiberius@ivdnt.org

the process of extending the corpus, which involves several different stages from translation to tokenization, lemmatization, and POS-tagging, to named entity and multiword expression/named entity annotation, and finally, word-sense disambiguation. We discuss some of the challenges encountered so far in these different phases with the corpus itself on the one hand, and the sense inventories on the other. We also describe the plans for future work on additional annotation layers within UniDive in order to further improve the ELEXIS-WSD corpus as a high-quality richly annotated manually curated dataset that is useful for NLP tasks such as word-sense disambiguation.

Keywords: semantic annotation, parallel corpus, senses, South Slavic languages, Slovene, Croatian, Serbian

1. Introduction

The ELEXIS-WSD Parallel Sense-Annotated Corpus is a multilingual parallel corpus compiled within the ELEXIS project (European Lexicographic Infrastructure; 2017–2020)[6] as part of a concerted effort to provide an open-source high-quality sense-annotated and linguistically informed dataset that can – among other things – be used for word-sense disambiguation tasks (hence the name ELEXIS-WSD). It is openly accessible as a dataset at the CLARIN.SI language repository (Martelli et al., 2023) – version 1.1 (the latest version at the time of writing this paper) is available under the Creative Commons BY-SA 4.0 license and covers 10 European languages: Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene, and Spanish.

The main incentive for the compilation of the ELEXIS-WSD corpus was the lack of high-quality manually curated sense-annotated datasets. During the ELEXIS project, existing datasets compiled for similar purposes were characterized by a number of drawbacks. First, the majority of them focused primarily on English (e.g., *SemCor* by Miller et al., 1993; datasets compiled within Senseval and SemEval shared tasks by Edmonds & Cotton, 2001; Snyder & Palmer, 2004; Navigli et al., 2007; Pradhan et al., 2007) or included only a limited number of instances in other languages (e.g., Agirre et al., 2010; Navigli et al., 2013; Moro & Navigli, 2015). Second, they relied on outdated or limited sense inventories or did not cover all part-of-speech categories. The endeavor within the ELEXIS project aimed to compile a resource that would link an annotated corpus with a high-quality sense inventory compiled through a lexicographic process (e.g. monolingual dictionaries).

[6] The ELEXIS Project Website: https://project.elex.is/

The compilation process and structure of the corpus are described in more detail by Martelli et al. (2021). Each language subcorpus of ELEXIS-WSD contains the same 2,024 sentences with a total of approximately 35,000 tokens per language (a total of approximately 345,000 tokens). For each language, there is a separate sense inventory file containing senses, their IDs and definitions. Each content word in the corpus (adjectives, adverbs, verbs, and nouns according to the Universal Dependencies annotation scheme) is annotated with its corresponding sense from the sense inventory.

The purpose of the paper is to present the on-going work on extending the ELEXIS-WSD corpus with new languages and/or additional annotation layers, an effort that is being carried out within the CA21167 COST Action (UniDive). In this paper, we focus on the progress and challenges presented by the subcorpora of three South Slavic languages: Croatian, Serbian, and Slovene. We first describe the general goals and plans of the ELEXIS-WSD extension task within UniDive (Section 2), then proceed with describing the current state of the three subcorpora and the tasks that have been carried out so far (Section 3). We describe the plans for future developments within UniDive and conclude the paper with lessons learned and open questions for future work (Section 4).

2. Extension of ELEXIS-WSD within UniDive

The extension of ELEXIS-WSD is one of the tasks within the CA21167 COST Action titled *Universality, Diversity and Idiosyncrasy in Language Technology*[7] (UniDive; 2022–2026), an international scientific network with the goal of reconciling language diversity with rapid progress in language technology. This provides the opportunity to extend the ELEXIS-WSD corpus in two ways: by adding other parallel subcorpora and thereby improving the corpus's coverage in terms of language diversity, and by adding additional annotation layers not present in the existing version, which will increase the usefulness of the corpus for a wide range of natural language processing tasks, as well as help develop and test out universal annotation guidelines and schemes across languages. The compilation process will be useful as a blueprint for potential similar datasets in the future, which will be crucial as evaluation benchmarks in the age of large language models.

Two elements need to be taken into account when adding a language to the ELEXIS-WSD corpus: the subcorpus (i.e. the set of 2,024 sentences)

[7] UniDive Website: https://unidive.lisn.upsaclay.fr/

and the sense inventory (i.e. the set of senses and definitions with which the content words in the subcorpus are annotated).

At the time of writing this paper, a total of 8 new languages are included in the extension process: Greek, Romanian, Georgian, Macedonian, Polish, Ukrainian, Croatian, and Serbian. The ELEXIS-WSD corpus was initially compiled by extracting English sentences and their translation equivalents in other languages from the WikiMatrix dataset of parallel sentences (Schwenk et al., 2021) according to several criteria such as sentence length and number of polysemous words in the English sentence (see Martelli et al., 2021). Any missing translations were then translated manually. Within UniDive, we use English as a pivot language and start by first translating the sentences using a machine translation system (depending on the language), then manually correcting the result. The reason for skipping the sentence extraction from WikiMatrix for new languages within UniDive is that in the first version of the corpus, even with high-resource languages (such as Spanish), a significant amount of parallel sentences were missing, while some of those extracted were erroneous, poorly translated or machine-translated. Since we deal mostly with less-resourced languages in UniDive, we decided to use machine translations as a starting point.

The sense inventory which will be used to annotate tokens in the subcorpus should be in a machine-readable format (e.g. XML, TSV) and should ideally cover the majority of the approximately 5,000 ADV (adverb), ADJ (adjective), NOUN (common noun) and VERB (verb) lemmas from the ELEXIS-WSD corpus, along with their Universal Part-of-speech (UPOS) tags, the division of each lexeme into senses, and sense definitions. For instance, if the English lemma *premiere* appears as a verb in the subcorpus, the sense inventory should include all the available senses for *premiere ~ VERB* even if only one of the senses appears in the corpus. An example is shown in Table 1 to demonstrate the structure of the sense inventory of the English subcorpus, which is derived from the Open English WordNet (McCrae et al., 2019).

Table 1: Structure of the English sense inventory in ELEXIS-WSD 1.1.

Lemma	UPOS	ID[8]	Definition
premiere	NOUN	...e1c-0	the first public performance of a play or movie
premiere	VERB	...e1d-0	be performed for the first time
premiere	VERB	...e1d-1	perform a work for the first time

[8] The IDs in Table 1 were shortened for demonstration purposes.

The sense inventory should be available under the Creative Commons BY-SA 4.0 license and should preferably be based on lexicographic data (monolingual dictionaries or other lexicographic databases); because dictionaries are frequently not open-source and difficult to obtain, other semantic resources such as WordNets and Wiktionaries are acceptable as a fall-back.

After the machine translation phase, the process of extending the corpus involves several different annotation stages (such as tokenization, lemmatization, and POS-tagging). We describe these in more detail in the following sections using the examples of the subcorpora of South Slavic languages.

3. South Slavic Languages in ELEXIS-WSD

Currently, ELEXIS-WSD contains subcorpora for 5 South Slavic languages at various stages of development. Bulgarian and Slovene were included in the first phase of the compilation, so the initial annotation layers are complete. Bulgarian is currently not being developed further within UniDive, while Slovene continues with additional annotation layers (see Section 3.3). Macedonian, Croatian, and Serbian are new languages and started from the beginning within UniDive. Macedonian is in its early stages of editing machine translations, while Croatian (Section 3.1) and particularly Serbian (Section 3.2) have already advanced into subsequent stages.

3.1. Croatian Subcorpus – ELEXIS-WSD-hr

The Croatian subcorpus of ELEXIS-WSD started with the automatic translation and manual validation of English sentences. The sentences were translated into Croatian twice by two different systems: *Google Translate* and *Hrvojka*,[9] a custom-made Croatian national machine-translation platform developed within the NLPT project for the benefits of public administration, the economic and academic communities. The Croatian UniDive team took parallel machine-translations and compared them to the English original to provide the final manually-curated Croatian version of each sentence.

The challenges with translation stemmed mostly from the nature of the sentences included in the corpus. The sentences are unrelated and disjointed as they are extracted from different Wikipedia texts. Their content is frequently very encyclopaedic and includes specific terms from

[9] *Hrvojka* is available as part of the Croatian National Language Technology Platform: https://hrvojka.gov.hr/

different scientific areas. In some cases, the sentences are very short and contain some unclear references (such as pronouns), for which gender and number sometimes had to be disambiguated. This was done by either consulting other subcorpora within ELEXIS-WSD or searching for the context of the original sentences on Wikipedia.

The translations have recently been finished and the subcorpus will proceed with the next annotation phases, which include tokenization, lemmatization, and part-of-speech tagging. This will be done using Universal Dependencies 2.15 tagging models available in *UDPipe 2*,[10] specifically the *croatian-set-ud-2.15-241121* model. The layers will then be manually corrected in the open-source INCEpTION annotation platform (Klie et al., 2018) before proceeding to more advanced annotations such as word-sense disambiguation (see Section 4).

As for the sense inventory that will be used to annotate senses in the subcorpus, ELEXIS-WSD-hr will use the *School Dictionary of Croatian* (*Školski rječnik hrvatskoga jezika*; Birtić et al., 2012),[11] a normative dictionary that contains approximately 31,000 entries and is intended for Croatian students in the final years of primary and secondary schools. Beside grammatical and morphological information on entries (part-of-speech, aspect, gender, etc.) and their sense definitions, the dictionary also includes multi-word expressions, synonyms, antonyms, and normative labels. However, additional work is required to prepare the sense inventory. While the dictionary is available in an XML format, the data is unstructured, with different types of data (from morphological information to definitions and multiword expressions) contained in a single string that is formatted according to lexicographic guidelines. Using rule-based approaches and regular expressions, the dictionary is currently being converted into the appropriate format (see Table 1). This is a work-in-progress that will require some additional manual verification. The initial phases of the process have been described by Runjaić & Čibej (2025). Any lemmas from ELEXIS-WSD-hr that are absent in the dictionary will be manually added to the sense inventory along with their sense divisions and definitions.

3.2. Serbian Subcorpus – ELEXIS-WSD-sr

Unlike Croatian, the sentences for the Serbian subcorpus (in Latin script) were translated only with Google Translate, then manually validated by 8 Serbian native speakers (Krstev et al., 2024a) in order to avoid literal

[10] UDPipe 2: https://lindat.mff.cuni.cz/services/udpipe/
[11] *The School Dictionary of Croatian* is also available in an online interface: https://rjecnik.hr/

or incorrect translations (particularly in the case of multi-word expressions) and to resolve issues already mentioned in Section 3.1 (resolution of ambiguous gender features in pronouns, specific terminology, etc.), a particular challenge for Serbian in this phase was the presence of a large number of foreign proper nouns and named entities, which according to Serbian orthographic rules should not be written in their original form, but transliterated phonetically. This was again validated by two people.

At the time of writing this paper, several annotation layers (tokenization, lemmatization, POS-tagging, as well as named entity recognition) have already been completed, while others are still in progress. Tokenization, lemmatization, and POS-tagging were first annotated automatically (Stanković et al., 2022) then checked by at least three evaluators. Named entities (NEs) were automatically annotated with the *sr_ner_tesla_j355* model (Ikonić Nešić et al., 2024) using a schema of 7 categories; an eighth category – PRODUCT – was introduced manually. In total, 2,294 NEs were manually verified: PERS (439), LOC (710), ORG (330), DEMO (301), ROLE (309), EVENT (60), WORK (75), PRODUCT (80). Named entities were manually linked with Wikidata using the INCEpTION platform (Klie et al., 2018). As a result, 1,949 NEs were linked with the existing Wikidata items, which leaves 345 unlinked NEs. The corpus was also automatically annotated with verbal and nominal MWEs (as elaborated in Krstev et al., 2024b), which was followed by manual validation. The classification of verbal MWEs followed the PARSEME guidelines,[12] while non-verbal MWEs will be categorized according to the extended PARSEME guidelines that are being developed within UniDive (more on this in Section 4).

Because there is no open-access digital descriptive dictionary of Serbian, the Serbian sense inventory for ELEXIS-WSD-sr is based on the *Serbian WordNet* (SrpWN; Stanković et al., 2018). However, SrpWN does not cover all the lemmas from the corpus and their senses, and needed to be expanded with the missing synsets. The inventory for the English subcorpus is derived from Open English WordNet, so its 13,703 synsets were aligned with SrpWN, resulting in 5,997 matches. A comparison with ELEXIS-WSD-en sense annotations highlighted 2,130 missing synsets that required urgent addition. The synonyms and definitions from Open English Wordnet were automatically translated using Google API and OpenAI and were later post-edited. For the rest of content words without definitions in the Serbian sense inventory, the entries were added by prompt engineering

[12] PARSEME Guidelines for MWE Annotation: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/

using Gemini (gemini-2.0-flash-exp) and ChatGPT (gpt-4), followed by post-editing. The Serbian inventory is currently being expanded with definitions of identified MWEs, mostly manually by consulting various resources.

The demo version of the current state of the annotated corpus is available as NIF (NLP Interchange Format) corpus.[13] Once completed, the ELEXIS-WSD-sr corpus will be the first sense-annotated corpus using the Serbian WordNet.

3.3. Slovene Subcorpus – ELEXIS-WSD-sl

As one of subcorpora from the original version of ELEXIS-WSD, ELEXIS-WSD-sl already has manually checked tokenization, lemmatization, part-of-speech tags, and sense annotations. Because no open-source Slovene dictionary was available at the time and the existing version of the Slovene WordNet – *sloWNet 3.1* (Fišer, 2015) – lacked a significant number of Slovene definitions and thus provided poor coverage for sense annotation, the sense inventory was compiled from scratch by co-locating the activity with ongoing efforts to develop the *Digital Dictionary Database of Slovene* (DDDS; Kosem et al., 2021). First, a frequency list of lemmas and their part-of-speech tags was extracted from ELEXIS-WSD-sl. For each lemma, a sense division was determined by lexicographers using both ELEXIS-WSD-sl and other Slovene corpora (such as the *Gigafida Corpus of Written Standard Slovene*; Krek et al., 2019). Individual occurrences of lemmas were then assigned appropriate senses from the DDDS, taking into account multi-word expressions (MWEs) as well – each individual token pertaining to a MWE was annotated with the sense of the entire expression.

While some information on MWEs can be extrapolated from sense annotations, the subcorpus does not yet explicitly contain multiword expression annotations. Within UniDive, however, all MWEs in ELEXIS-WSD-sl will be annotated with categories proposed first in the PARSEME COST Action for verbal MWEs (such as light-verb constructions and verbal idioms; see e.g. Savary et al., 2023) and further developed within UniDive for non-verbal MWEs (nominal, adjectival, adverbial, and functional MWEs; an initial taxonomy was proposed by Ramisch (2023)).

Another annotation layer concerns dependency relations according to the Universal Dependencies annotation scheme (for a description of the plans to add UD syntax annotations to ELEXIS-WSD, see Tiberius et al.

[13] The ELEXIS-WSD-sr corpus as NIF is available at endpoint: http://fuseki.jerteh.rs/#/dataset/sr-ELEXIS-UNIDIVE-demo/query

(2024); the plan is to have every language annotated with at least automatically assigned (or, ideally, manually validated) dependency relations). The sentences from ELEXIS-WSD-sl were also included in the *SUK Training Corpus of Written Slovene* (Arhar Holdt et al., 2024), where they were manually annotated with dependency relations. However, in the process, some changes occurred in morphosyntactic tagging as well, so the mapping process between the SUK and ELEXIS-WSD-sl versions is not entirely straightforward. The discrepancies are currently being resolved.

4. Future Plans

In addition to the annotation layers already mentioned (lemmatization, morphosyntactic tags, sense annotations, UD dependency relations, and MWE annotations according to the PARSEME/UniDive categorization), the finalized versions of the subcorpora in ELEXIS-WSD will also contain (at least) two additional annotation layers.

The first layer concerns named entities (NEs). For the English subcorpus, NEs have already been annotated, but the annotations have not yet been implemented in version 1.1 as they were only meant as a reference for sense annotation. In addition, only spans were annotated. In the Serbian subcorpus, however, NEs have been annotated much more systematically (Krstev et al., 2024a) using a set of 8 classes. There is still the open question of the set of categories used for annotation and whether the set is universal enough to be applied to each language within ELEXIS-WSD (see Krstev et al., 2024b). Unlike the PARSEME guidelines for categorizing MWEs, NEs currently have no universal guidelines within UniDive yet (an example of annotating a corpus with both MWEs and NEs is provided by e.g. Candito et al. 2020). Once the annotation scheme is complete, the NE annotations will be propagated across all subcorpora. In addition, the annotated NEs will be further enriched by links to their corresponding WikiData instances.

The second layer is an additional semantic layer alongside the existing sense annotations that link individual tokens to senses in the sense inventory. Each sense annotation will also be annotated with a *supersense*, a more coarse-grained semantic category denoting broad concepts such as ANIMAL, ARTIFACT, FEELING, GROUP, and PLANT. The set of supersenses are used to more broadly categorize groups of semantically similar synsets in WordNets (Jurafsky & Martin, 2025). There are a total of 44 supersense categories (26 for nouns, 15 for verbs, 2 for adjectives, and 1 for adverbs). Because a number of subcorpora in ELEXIS-WSD already use different WordNets as sense inventories, the propagation of supersense

annotations from synsets onto individual sense-annotated tokens is a trivial task. Propagation to other subcorpora which use non-WordNet sense inventories will require some more work, but to an extent, the process can be automated with large language models to reduce the amount of manual validation required.

5. Conclusion

We have briefly presented the on-going effort of extending the ELEXIS-WSD Parallel Sense-Annotated Corpus with two new South Slavic languages (Croatian and Serbian), as well as additional work on the existing Slovene subcorpus. We described some of the challenges encountered in different annotation phases of the corpus, as well as different solutions to providing an open-access sense inventory for corpus annotation.

Just like version 1.1, all the data produced will be openly accessible at the CLARIN.SI language repository under the Creative Commons BY-SA 4.0 licence. The sense inventories will also be uploaded to WikiBase[14] and will be available for querying at a SPARQL endpoint.

The work within UniDive will result in a richly annotated multi-layered corpus that can be useful for NLP tasks or contrastive linguistic analyses. The experience of compiling the dataset will be invaluable in the future as similar resources are developed as benchmark datasets to evaluate the performance of large language models for complex semantic tasks.

Acknowledgment

The work presented in the paper was supported by the *COST Action CA21167 – Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive). The authors also acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – *Language Resources and Technologies for Slovene*). A sincere word of gratitude also goes to the anonymous reviewers for their insightful comments.

[14] Wikibase: https://wikiba.se/

References

[1] Agirre, Eneko, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen and Roxanne Segers. "SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain". *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden. Association for Computational Linguistics. (2010): 75–80. https://aclanthology.org/S10-1013/

[2] Arhar Holdt, Špela, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, Anja Zajc, 2024, Training corpus SUK 1.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1959.

[3] Birtić, Matea, Goranka Blagus Bartolec, Lana Hudeček, Ljiljana Jojić, Barbara Kovačević, Kristian Lewis, Ivana Matas Ivanković, Milica Mihaljević, Irena Miloš, Ermina Ramadanović and Domagoj Vidović. "*Školski rječnik hrvatskoga jezika*". Školska knjiga, Zagreb, Croatia. (2012)

[4] Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, Silvio Ricardo Cordeiro. "A French corpus annotated for multiword expressions and named entities". *Journal of Language Modelling, 8(2)*. (2020): 415–479. https://doi.org/10.15398/jlm.v8i2.265

[5] Edmonds, Philip and Scott Cotton. "SENSEVAL-2: Overview". *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Toulouse, France. Association for Computational Linguistics. (2001): 1–5. https://aclanthology.org/S01-1001/

[6] Fišer, Darja. "Semantic lexicon of Slovene sloWNet 3.1". Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2015) http://hdl.handle.net/11356/1026

[7] Ikonić Nešić, Milica, Saša Petalinkar, Mihailo Škorić, Ranka Stanković. "BERT Downstream Task Analysis: Named Entity Recognition in Serbian". In: Trajanović, M., Filipović, N., Zdravković, M. (eds) Disruptive Information Technologies for a Smart Society. ICIST 2024. Lecture Notes in Networks and Systems, vol 860. Springer, pp 333—347 Cham. (2024) https://doi.org/10.1007/978-3-031-71419-1_29

[8] Jurafsky, Daniel and James H. Martin. "Word Senses and WordNet. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models". (2025). https://web.stanford.edu/~jurafsky/slp3/G.pdf

[9] Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho and Iryna Gurevych. "The INCEpTION Platform: Machine-Assisted and

Knowledge-Oriented Interactive Annotation". *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA (2018): 5–9. https://aclanthology.org/C18-2002/

[10] Kosem, Iztok, Simon Krek and Polona Gantar. "Semantic data should no longer exist in isolation: the digital dictionary database of Slovenian." *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion*. Komotini: SynMorPhoSe Lab, Democritus University of Thrace. (2021): 81–83. https://elex.is/wp-content/uploads/2021/09/Semantic-Data-should-no-longer-exist-in-isolation-the-Digital-Dictionary-Database-of-Slovenian_Kosem-Krek-Gantar_EURALEX2020.pdf

[11] Krek, Simon, Carole Tiberius, Kaja Dobrovoljc, Jaka Čibej, Polona Gantar, Jelena Kallas, Kristina Koppel, Svetla Peneva Koeva, Veronika Lipp, László Simon: "The ELEXIS Parallel Sense-annotated Corpus". *UniDive 1st General Meeting*. Paris, France. (2023). https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:2023-saclay:abstracts:45_krek_et_al_the_elexis_parallel_sense_annotated_corpus.pdf

[12] Krek, Simon, Tomaž Erjavec, Andraž Repar, Jaka Čibej, Špela Arhar Holdt, Polona Gantar, Iztok Kosem, Marko Robnik-Šikonja, Nikola Ljubešić, Kaja Dobrovoljc, Cyprian Laskowski, Miha Grčar, Peter Holozan, Simon Šuster, Vojko Gorjanc, Marko Stabej, Nataša Logar. "Corpus of Written Standard Slovene Gigafida 2.0". *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042. (2019) http://hdl.handle.net/11356/1320.

[13] Krstev, Cvetana, Aleksandra Marković, Ranka Stanković. "Annotation of MWEs and NEs in the Serbian extension of ELEXIS-WSD: comparisons, solutions and open questions." *Unidive 2nd General Meeting*. Naples, Italy. (2024b) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:2nd_unidive_general_meeting:17_annotation_of_mwes_and_nes_in_.pdf

[14] Krstev, Cvetana, Aleksandra Marković, Ranka Stanković, Milica Ikonić Nešić. "Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository." *Unidive 3rd General Meeting*. Budapest, Hungary. (2025) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:32_progress_in_sr_elexis_seman.pdf

[15] Krstev, Cvetana, Bojana Djordjević, Sanja Antonić, Nevena Ivković-Berček, Zorica Zorica, Vesna Crnogorac, and Ljiljana Macura. "Cooperative Work in Further Development of Serbian WordNet". *INFOteka, IX(12)*. (2008): 59–78.

[16] Krstev, Cvetana, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov. "Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities". *Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)*, SIGLEX, ACL,

UniDive CA21167. eds.: Bhatia, Archna et al. virtual presentation, May 25, 2024. (2024a): 74-84. https://aclanthology.org/2024.mwe-1.15.pdf

[17] Martelli, Federico, 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, Rafel 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, Sia Kolkovska: "Parallel sense-annotated corpus ELEXIS-WSD 1.1". *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042. (2023). http://hdl.handle.net/11356/1842

[18] Martelli, Federico, 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. "Designing the ELEXIS Parallel Sense-annotated Dataset in 10 European Languages." *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*. (2021): 377–395. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_22_pp377-395.pdf

[19] McCrae, John P., Alexandre Rademaker, Francis Bond, Ewa Rudnicka and Christiane Fellbaum. "English WordNet 2019 – An Open-Source WordNet for English". *Proceedings of the 10th Global WordNet Conference – GWC 2019, Wrocław*. (2019): 245–252. https://aclanthology.org/2019.gwc-1.31/

[20] Miller, George A., Claudia Leacock, Randee Tengi, Ross T. Bunker. "A semantic concordance." *Human Language Technology: Proceedings of a Workshop Held at Plainsboro*, New Jersey, March 21-24, 1993. (1993) https://aclanthology.org/H93-1061/

[21] Moro, Andrea and Roberto Navigli. "SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking". *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado. Association for Computational Linguistics. (2015): 288–297. https://aclanthology.org/S15-2049/

[22] Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. "SemEval-2007 Task 07: Coarse-Grained English All-Words Task". *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic. Association for Computational Linguistics. (2007): 30–35. https://aclanthology.org/S07-1006/

[23] Navigli, Roberto, David Jurgens, and Daniele Vannella. "SemEval-2013 Task 12: Multilingual Word Sense Disambiguation". *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA. Association for Computational Linguistics. (2013): 222–231. https://aclanthology.org/S13-2040/

[24] Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer. "SemEval-2007 Task-17: English Lexical Sample, SRL and All Words". *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic. Association for Computational Linguistics. (2007): 87–92. https://aclanthology.org/S07-1016/

[25] Ramisch, Carlos. "A taxonomy proposal for multiword expressions". *2nd UniDive General Meeting*. Naples, Italy. (2024) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:2nd_unidive_general_meeting:23_a_taxonomy_proposal_for_multiw.pdf

[26] Runjaić, Siniša and Jaka Čibej. "Conversion of the School Dictionary of Croatian into a Sense Inventory for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Case Study". *UniDive 3rd General Meeting*. Budapest, Hungary. (2025) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:20_conversion_of_the_school_di.pdf

[27] Savary, Agata, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, Joakim Nivre. "PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions". *Northern European Journal of Language Technology*, Vol. 9, No. 1. (2023). https://nejlt.ep.liu.se/article/view/4453

[28] Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia". *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. (2021): 1351–1361. https://aclanthology.org/2021.eacl-main.115/

[29] Snyder, Benjamin and Martha Palmer. "The English all-words task". *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain. Association for Computational Linguistics. (2004): 41–43. https://aclanthology.org/W04-0811/

[30] Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. "Resource-based WordNet Augmentation and Enrichment". *Proceedings of the Third International Conference Computational Linguistics in Bulgaria* (CLIB 2018), pages 104–114, Sofia, Bulgaria. Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences. (2018)

[31] Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel Bidirectionally Pretrained Taggers as Feature Generators" Applied Sciences 12, no. 10: 5028. https://doi.org/10.3390/app12105028

[32] Tiberius, Carole, Jaka Čibej, Jelena Kallas, Kertu Saul, Kadri Muischnek, Simon Krek. "UD Syntax for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Pilot Study". *UniDive 2nd General Meeting*. Naples, Italy. (2024) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:2nd_unidive_general_meeting:31_ud_syntax_for_the_elexis_paral.pdf

ELEXIS-WSD paralelni semantički anotirani korpus i južnoslovenski jezici: potkorpusi za hrvatski, srpski i slovenački

Jaka Čibej, Ranka Stanković, Ana Ostroški Anić, Simon Krek, Carole Tiberius

Sažetak

Otvoreni paralelni semantički anotiran korpus ELEXIS-WSD razvijen je u okviru projekta ELEXIS i u verziji 1.1 sadrži 2.024 rečenice za svaki od 10 jezika: bugarski, danski, engleski, španski, estonski, mađarski, italijanski, holandski, portugalski i slovenački. U rečenicama, svakoj reči koja nose značenje (imenice, pridevi, glagoli i prilozi) dodeljeno je odgovarajuće značenje iz jednog od 10 otvorenih repozitorijuma značenja koji sadrže definicije. U kontekstu UniDive COST akcije (CA21167), korpus se proširuje sa nekoliko novih jezika, uključujući južnoslovenske jezike. U ovom radu fokusiramo se na tri potkorpusa: hrvatski, srpski i slovenački. Ukratko opisujemo strukturu i namenu ELEXIS-WSD korpusa, a zatim nastavljamo sa opisom procesa proširenja korpusa, koji uključuje nekoliko različitih faza – od prevođenja do tokenizacije, lematizacije i obeležavanja vrsta reči (POS-tagging), zatim anotacije imenovanih entiteta i polileksemskih izraza/imenovanih entiteta, i konačno, razrešavanja višeznačnosti reči. U radu razmatramo neke od izazova sa kojima smo se do sada susreli u različitim fazama, kako u pripremi korpusa, tako i u pripremi repozitorijuma značenja. Takođe opisujemo planove za budući rad na dodatnim slojevima anotacije u okviru UniDive projekta, sa ciljem daljeg poboljšanja ELEXIS-WSD korpusa kao visokokvalitetnog, bogato anotiranog i ručno verifikovanog skupa podataka koji je koristan za NLP zadatke kao što je razrešavanje višeznačnosti reči.

Ključne reči: semantička anotacija, paralelni korpus, značenja, južnoslovenski jezici, slovenački, hrvatski, srpski