
ELEXIS-WSD 2.0: Expanding the Lexicon–Corpus Interface with New Languages and Annotation Layers

Jaka Čibej¹[0000-0002-3037-6848], Simon Krek¹[0000-0001-8965-6863], Carole
Tiberius²[0000-0002-7860-5427], Irina Lobzhanidze³[0000-0002-3040-1847],
Verginica Barbu Mititelu⁴[0000-0003-1945-2587], Jelena
Kallas⁵[0000-0001-8614-0140], Kertu Saul⁵[0009-0002-0964-0015], Kadri
Muischnek⁶[0000-0002-3920-6618], Karoliina Jõgi⁶, Ranka
Stanković⁷[0000-0001-5123-6273], Cvetana Krstev⁷[0000-0003-3328-9392],
Aleksandra Marković⁷[0000-0003-2985-3042], Milica Ikonić
Nešić⁷[0000-0002-0835-8889], Ana Ostroški Anić⁸[0000-0001-9999-0750],
Bartłomiej Alberski⁹[0000-0002-9099-6627], Vladimir
Cvetkoski¹⁰[0000-0002-8570-1062], Voula Giouli¹¹[0000-0002-1501-9754],
Rusudan Makhachashvili¹²[0000-0002-4806-6434], Olha
Kanishcheva¹³[0000-0002-9035-1765]

¹ University of Ljubljana, Slovenia, {jaka.cibej|simon.krek}@ff.uni-lj.si

² Dutch Language Institute, The Netherlands, carole.tiberius@ivdnt.org

³ Iliia State University, Georgia, irina_lobzhanidze@iliauni.edu.ge

⁴ Research Institute for Artificial Intelligence, Romania, vergi@racai.ro

⁵ Institute of the Estonian Language, Estonia, {jelena.kallas|kertu.saul}@eki.ee

⁶ University of Tartu, Estonia, {kadri.muischnek|karoliina.jogi}@ut.ee

⁷ Language Resources and Technologies Society – JeRTeh, Serbia,
{cvetana|ranka|milicain}@jerteh.rs, aleksandra.markovic@isj.sanu.ac.rs

⁸ Institute for the Croatian Language, Croatia, aostrosk@ihjj.hr

⁹ Wrocław University of Science and Technology, Poland,
bartlomiej.alberski@pwr.edu.pl

¹⁰ Ss. Cyril and Methodius University in Skopje, North Macedonia,
cvetkoski@ff.ukim.edu.mk

¹¹ Aristotle University of Thessaloniki, Greece, pgiouli@del.auth.gr

¹² Borys Grinchenko Kyiv Metropolitan University, Ukraine,
r.makhachashvili@kubg.edu.ua

¹³ Heidelberg University, Germany, kanichshevaolga@gmail.com

Abstract

The ELEXIS-WSD Parallel Sense-Annotated Corpus is a multilingual resource designed to support word sense disambiguation, lexical-semantic research, and cross-lingual studies by linking corpus occurrences directly to lexical resources through manual sense annotation. Within the UniDive COST Action, WG2 (Task 2.2) focused on extending and enriching this resource, resulting in the release of ELEXIS-WSD 2.0 in April 2026 (Čibej et al., 2026). The new version significantly expands the corpus by introducing eight additional languages: Romanian, Greek, Ukrainian, Polish, Macedonian, Georgian, Croatian, and Serbian, while further enriching existing subcorpora for Slovene, Estonian, and Dutch.

The extension process involved multiple annotation layers. Newly added language subcorpora were compiled through automatic translation of 2,024 English source sentences followed by manual correction and validation. The corrected texts were tokenized, lemmatized, morphosyntactically annotated, and syntactically parsed, with manual corrections performed where resources permitted. Additional semantic layers include the annotation of multiword expressions according to the PARSEME 2.0 guidelines and named entities linked to Wikidata. The Serbian team contributed with named entity annotation and linking, as well as the first sense-annotated subcorpus through a semi-automatic annotation workflow.

As a result, ELEXIS-WSD 2.0 now comprises approximately 38,000 sentences and 620,000 tokens, representing an increase of about 80% compared to the pre-UniDive version. The corpus is openly available through the CLARIN.SI repository and provides a valuable infrastructure for multilingual lexical-semantic research. Future work will focus on completing remaining annotation layers, harmonizing annotations across languages, expanding sense inventories, and exploring the use of large language models for automatic sense generation and linking.

Keywords: word sense disambiguation, parallel corpus, semantic annotation, named entities, multiword expressions, lexical resources, ELEXIS-WSD

1. Introduction

Within the COST Action CA21167 – *Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive) (Savary et al., 2024), Task 2.2 (*Design of a Lexicon-Corpus Interface*) of Working Group 2 focuses on the development and enhancement of the *ELEXIS-WSD Parallel Sense-Annotated Corpus* (*ELEXIS-WSD*). This multilingual resource consists of parallel subcorpora containing translations of the same 2,024 sentences in multiple languages. The

original *ELEXIS-WSD* corpus was compiled by selecting English sentences and their multilingual counterparts from the WikiMatrix parallel corpus (Schwenk et al., 2021), based on criteria such as sentence length and the presence of polysemous words (Martelli et al., 2021). Missing translations were subsequently produced manually. In the UniDive extension, English was retained as the pivot language; however, due to data gap and quality issues, instead of extracting additional sentence pairs from WikiMatrix, new language subcorpora were generated through machine translation followed by manual revision. The *ELEXIS-WSD* corpus comprised manually validated tokenization, lemmatization, morphosyntactic annotation, named entity recognition and linking, multiword expression annotation, and, most importantly, manually assigned word-sense annotations derived from lexical resources such as dictionaries and wordnets. By explicitly linking corpus occurrences to lexical entries, *ELEXIS-WSD* serves as a bridge between corpora and lexicons, enabling researchers to study lexical meaning in authentic multilingual contexts.

Resources of this type are particularly valuable for research in word sense disambiguation, lexical semantics, multilingual natural language processing, and cross-lingual language comparison. The availability of parallel sense-annotated data facilitates the investigation of semantic correspondences and divergences across languages, supports the evaluation and training of WSD systems, and provides a reliable benchmark for the development of language technologies. Furthermore, such resources are becoming increasingly important in the era of large language models, as they offer high-quality, manually curated semantic data that can be used for evaluation, fine-tuning, and the development of semantically informed AI systems.

Prior to the UniDive initiative, the latest release of the corpus was version 1.1 (Martelli et al., 2023), which included subcorpora for ten languages: Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene, and Spanish. Within UniDive, significant efforts were devoted to extending both the linguistic coverage and the annotation depth of the resource.

In this paper, we present the outcomes of Task 2.2, which culminated in the publication of *ELEXIS-WSD* version 2.0 in April 2026.¹ The main objectives were twofold: (1) to broaden the multilingual scope of the corpus through the inclusion of additional languages,

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

and (2) to enrich the corpus with new annotation layers, thereby increasing its usefulness for linguistic research, lexicography, digital humanities, and advanced NLP applications (Čibej et al., 2026).

2. Linguistic Annotation Layers

Task 2.2 brought together 11 language teams working collaboratively on the expansion and enrichment of the ELEXIS-WSD corpus. The effort resulted in the addition of eight new language subcorpora: Romanian, Greek, Ukrainian, Polish, Macedonian, Georgian, Croatian, and Serbian, substantially increasing the multilingual coverage of the resource. In parallel, three teams responsible for existing subcorpora (Slovene, Estonian, and Dutch) focused on introducing new annotation layers and enhancing previously available linguistic information.

Given the diversity of participating languages, available resources, and team sizes, the extension process did not proceed uniformly across all language groups. Some teams joined the initiative earlier and were able to dedicate larger numbers of annotators and experts, while others faced limitations related to human resources, available language technologies, or lexical resources. Consequently, the newly added and existing subcorpora reached different stages of completion by the time version 2.0 was released. While some language teams completed the full annotation workflow, others focused on specific layers and will continue their work in subsequent releases.

To accommodate these differences while maintaining a common methodological framework, the extension activities were organized into a series of annotation layers, each representing a distinct stage of corpus enrichment. These layers include translation validation, tokenization, lemmatization, morphosyntactic annotation, syntactic parsing, multiword expression annotation, named entity annotation and linking, and, where available, sense annotation. The following subsections describe each stage of the extension process in detail and summarize the progress achieved across participating languages.

2.1. Translation

For each newly introduced language, the extension process began with a set of 2,024 source sentences taken from the English subcorpus, which serves as the pivot language of the ELEXIS-WSD

corpus. Initial translations were generated automatically using one of several machine translation systems, including *Google Translate*, *Gemini*, *DeepL*, and *GPT-4o*, depending on language coverage, availability, and the preferences of individual language teams. The use of automatic translation significantly accelerated the compilation of new language subcorpora while ensuring consistency with the original English source material.

However, because the corpus is intended for high-quality linguistic and semantic annotation, all automatically generated translations underwent careful manual revision. Annotators systematically compared each translated sentence with the English original, correcting lexical, grammatical, and stylistic inaccuracies. Particular attention was devoted to polysemous and ambiguous lexical items, whose incorrect translation could compromise the semantic alignment between languages and negatively affect subsequent annotation layers.

The revision process also considered factors relevant to downstream linguistic processing. Translators ensured that sentence structures, token boundaries, multiword expressions, and named entities were rendered in a way that would facilitate subsequent tokenization, lemmatization, morphosyntactic annotation, and sense annotation. All corrections were performed collaboratively in Google Sheets, which provided a convenient environment for distributed annotation, version control, quality assurance, and cross-checking among team members. This combination of machine translation and expert manual validation ensured both efficiency and linguistic reliability in the creation of the new multilingual subcorpora.

2.2. Tokenization

Following manual revision, the translated texts were automatically tokenized, primarily using *UDPipe* (Straka et al., 2016; Straka, 2018), although some language teams employed language-specific tools where appropriate. The tokenization process relied on existing Universal Dependencies models, including *romanian-rrt-ud-2.12-230717*, *polish-pdb-ud-2.15-241121*, *georgian-glc-ud-2.15-241121*, *croatian-set-ud-2.15-241121*, and *greek-gdt-ud-2.17-251125*. Due to the lack of suitable processing tools, the Macedonian subcorpus was annotated using Gemini 3.1 Pro through grammar-oriented prompting. The Serbian subcorpus was tokenized using language-specific models and procedures described in (Stanković et al., 2022). At the

time of the ELEXIS-WSD 2.0 release, the Ukrainian subcorpus had not yet undergone tokenization, so only the translations of the sentences are included. The tokenized version will be included in future versions of the resource.

The use of automatic tokenization ensured methodological consistency across languages while reducing the amount of manual effort required for subsequent annotation stages. At the same time, the flexibility to employ language-specific tools and workflows allowed individual teams to accommodate the particular linguistic and technological characteristics of their languages.

The automatically generated tokenizations were subsequently reviewed and corrected manually using Google Sheets and a modified CoNLL-U representation. Particular attention was devoted to the identification and correction of subtokens and multiword forms, as these are crucial for the later stages of semantic annotation, especially word-sense assignment. The tokenization strategy was therefore aligned with the requirements of the sense annotation layer, ensuring that each lexical unit could be linked to an appropriate sense inventory entry.

Manual subtokenization was especially important in cases where a single orthographic word corresponds to multiple semantic or grammatical units. For example, certain past-tense verb forms in Polish contain clitic elements that can be separated during tokenization and morphosyntactic annotation. The form *chciałby* ‘would like to’ was subtokenized into *chciał* ‘like’ and *by* ‘would’. The first segment represents the lexical part of the verb, while the second corresponds to the auxiliary verb marking the conditional. Such subtokenization enables more precise lemmatization, morphosyntactic analysis, and lexical-semantic annotation, as individual components can be processed and linked to lexical resources independently. This approach is particularly useful in multilingual corpora and word-sense annotation tasks, where a finer-grained representation of linguistic units improves consistency across languages.

To ensure annotation quality and structural coherence, the manually corrected data underwent automatic consistency checks. These validation procedures ensured that token and subtoken segmentation accurately reflected sentence boundaries and that all annotated units were properly aligned with the source text.

2.3. Lemmatization and Morphosyntactic Annotation

Following manual tokenization, the subcorpora underwent automatic linguistic annotation, including lemmatization, morphosyntactic tagging, and syntactic parsing. Most language teams employed *UDPipe* as the primary processing framework, while language-specific solutions were used where appropriate. For Macedonian, linguistic annotation was performed using Gemini 3.1 Pro through grammar-oriented prompting. The Serbian subcorpus was tokenized using language-specific models and procedures described in (Stanković et al., 2020; Stanković et al., 2022).

The resulting annotations were stored in the CoNLL-U format and subsequently converted to TSV3, a format compatible with the *INCEpTION* annotation platform (Klie et al., 2018). This conversion enabled collaborative annotation and facilitated the integration of additional annotation layers within a unified environment.

To support distributed work between participating institutions, dedicated user accounts were created for each language team. In cases where multiple annotators contributed to the same language subcorpus, the annotation process was coordinated by a language leader responsible for task allocation, quality control, and consistency checking. This collaborative workflow ensured methodological coherence while accommodating the specific requirements and expertise of individual language teams.

The primary objective of lemmatization and morphosyntactic annotation was not only to enrich the corpus with standard linguistic information, but also to prepare the data for subsequent lexical-semantic annotation. Particular attention was therefore paid to the accurate identification of lexical units and their grammatical properties, as these constitute essential prerequisites for assigning word senses from the corresponding sense inventories. By establishing a reliable layer of morphosyntactic information, the corpus provides a solid foundation for word-sense annotation, lexical-semantic research, and cross-lingual comparison.

Figure 1 presents the distribution of major part-of-speech categories across the language subcorpora included in ELEXIS-WSD 2.0, with focus on content words. Nouns constitute the largest category in all languages, followed by verbs and adjectives, reflecting the lexical-semantic focus of the corpus. Despite typological differences among the languages, the overall POS distributions are relatively sim-

ilar, indicating a high degree of comparability between the parallel subcorpora. Minor variations can be observed in the proportions of adpositions, proper nouns, and other categories, which are largely attributable to language-specific grammatical characteristics and translation choices. Overall, the results confirm the structural consistency of the corpus and its suitability for multilingual lexical-semantic and cross-lingual studies.

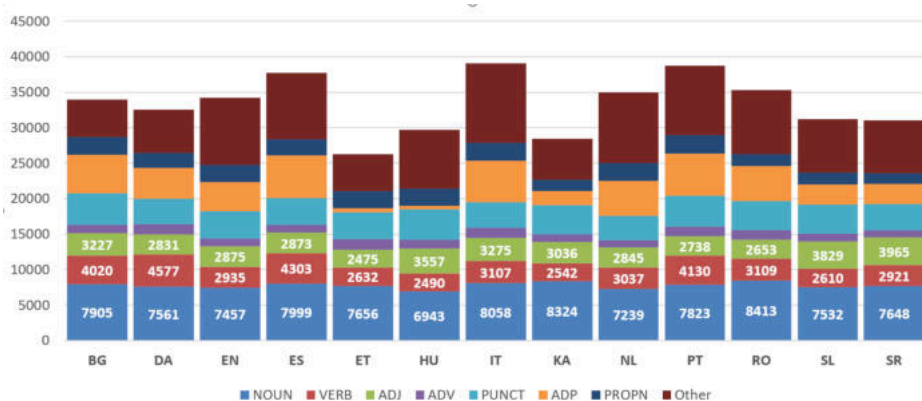


Figure 1: Universal POS tags in ELEXIS 2.0.

2.4. Syntactic Parsing

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.

3. Semantic Annotation Layers

3.1. Multiword Expressions

An additional annotation layer focused on the identification and annotation of multiword expressions (MWEs) following the *PARSEME 2.0* guidelines,² which define 32 categories of MWEs and

² PARSEME 2.0 Guidelines: <https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/>

provide a harmonized framework for multilingual annotation (Savary et al., 2026). These guidelines were developed within UniDive WG1 *Corpus Annotation* and have been applied across several corpora participating in the project, facilitating interoperability and cross-lingual comparison of MWE annotations.

Since the final version of the guidelines became available only in mid-2025, the MWE annotation effort started later than other annotation activities. Consequently, the level of completion differs across language teams. At the time of the ELEXIS-WSD 2.0 release, comprehensive MWE annotations were available for the Slovene and Serbian subcorpora. Annotation work is currently ongoing for the Estonian, Dutch, Georgian, and Romanian subcorpora and is expected to be completed by the end of the UniDive project. The inclusion of MWE annotations enriches the corpus with an additional semantic layer, supporting research on phraseology, lexical semantics, machine translation, and multilingual language technologies.

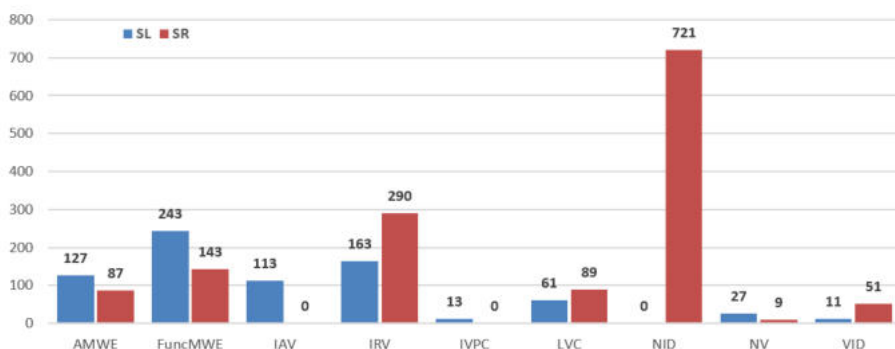


Figure 2: Serbian and Slovenian MWE annotation by main MWE types

To enhance readability, we have grouped some fine grained annotation categories:

- adjectival and adverbial MWEs (AMWEs), including adjectival and adverbial idioms, with separate subcategories for those derived from VMWEs: AdvID, AMWE, AV.IAV, AV.LVC.cause.
- functional MWEs (FuncMWEs), including multiword determiners, adpositions, conjunctions, and interjections: AdpID, ConjID, DetID, IntjID, PartID, PronID
- light verb constructions (LVCs) with two subcategories: LVC (LVC.cause, LVC.full)

- deverbal nominal MWE (NV) with subcategories corresponding to the categories of VMWEs from which the nominal MWE can be derived: NV.IAV, NV.LVC.full, NV.LVCfull, NV.VID

3.2. Named Entities

Named entities were already annotated in the English sub-corpus of the pre-UniDive version of *ELEXIS-WSD*; however, these annotations were limited to span identification and did not include semantic categorization or links to external knowledge bases. As a result, although entity mentions could be localized in the text, the annotations provided only limited support for semantic analysis and cross-lingual comparison.

Within UniDive, the Serbian language team substantially extended this annotation layer by introducing a fine-grained named entity annotation scheme consisting of eight semantic categories (Krstev et al., 2025). The annotation process resulted in the identification of more than 2,350 named entity mentions, covering persons (PER), locations (LOC), organizations (ORG), events (EVENT), works (WORK), products (PRODUCT), roles (ROLE), demonyms (DEMO), and other relevant entity types (OTHER), as shown in Figure 3. In addition to category assignment, entity mentions were linked to corresponding Wikidata entries whenever suitable candidates were available. This linking process transformed the corpus from a purely linguistic resource into a semantically enriched dataset connected to the Linked Open Data ecosystem.

The introduction of named entity linking provides several important benefits. First, it enables the normalization of different surface forms referring to the same real-world entity. Second, it facilitates interoperability with external resources such as Wikidata and other linked-data repositories. Third, it supports more advanced semantic analyses, including entity-centric search, knowledge graph construction, relation extraction, and cross-lingual entity alignment. The resulting annotations also create opportunities for studying how the same entities are represented across different languages and translations.

In future work, the Serbian annotations will serve as the basis for semi-automatic projection and harmonization of named entities in the remaining language subcorpora. By combining annotation transfer techniques with manual validation, the project aims to establish a

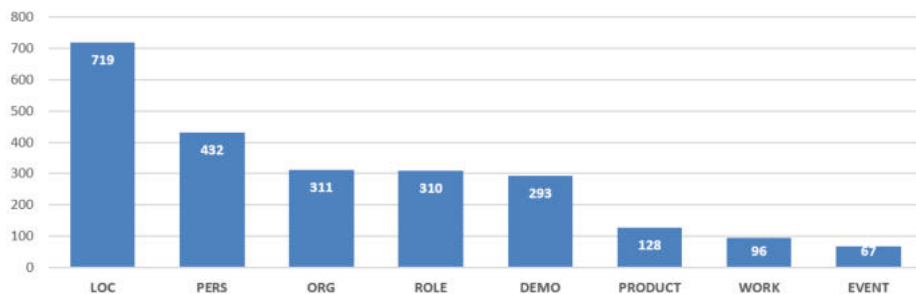


Figure 3: Named entities in sr-ELEXIS 2.0.

consistent multilingual layer of categorized and linked named entities across the entire *ELEXIS-WSD* corpus, further increasing its value for lexical-semantic research, multilingual NLP, and knowledge-based language technologies.

3.3. 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,³ 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 6.).

A crucial prerequisite for sense annotation is the availability of a machine-readable sense inventory, i.e. a lexical resource (such as a dictionary or wordnet) containing lexical units, their sense distinctions, and corresponding definitions. A sense inventory serves as the reference framework for assigning senses to lexical units in context. The preparation of such inventories proved to be one of the most demanding tasks within UniDive WG2. Many participating languages either lacked suitable open-access lexical resources or relied on resources that were incomplete, outdated, or not readily available in a format suitable for corpus annotation. Consequently, the creation, adaptation, and validation of sense inventories required substantial linguistic expertise and time investment. More details on sense inventories in ELEXIS 2.0 are provided in Section 4..

³ A sense inventory is a lexical resource (e.g., dictionary, WordNet) containing lexemes/lexical units, their sense divisions, and their corresponding definitions.

Among the newly introduced language subcorpora, Serbian is currently the only subcorpus that has progressed to the word-sense annotation stage. This achievement was made possible through the development of a hybrid semi-automatic annotation workflow that combines lexical resources, large language models, and expert validation (Stanković et al., 2026). The methodology was designed to reduce the considerable manual effort traditionally associated with word-sense annotation while maintaining a high level of annotation quality and consistency.

The Serbian team addressed this challenge by developing a dedicated sense inventory and a semi-automatic annotation pipeline, enabling the creation of the first sense-annotated subcorpus among the newly added languages. The experience gained through this process provides valuable methodological insights for future extensions of ELEXIS-WSD and demonstrates the potential of combining traditional lexical resources with modern AI-assisted approaches. The remaining language teams are expected to undertake sense annotation in future project phases, once the necessary lexical resources and annotation infrastructure become available (see Section 6.).

4. Sense Inventory

A key component of the *ELEXIS-WSD* framework is the use of language-specific sense inventories, which provide the semantic foundation for word-sense annotation. In previous versions of the corpus, sense annotation was performed using the *LexTag* platform,⁴ where each content word, namely nouns, verbs, adjectives, and adverbs, was manually assigned the most appropriate sense from a selected lexical resource or, when necessary, from BabelNet (Martelli et al., 2023). This approach established a direct connection between corpus occurrences and lexicographic resources, enabling the corpus to function as a lexicon-corpus interface rather than merely a collection of annotated texts.

The sense inventories employed in *ELEXIS-WSD* originate from lexical resources available for individual languages. These inventories differ in size, structure, and lexicographic tradition, ranging from explanatory dictionaries to wordnets and integrated lexical databases. During the annotation process, annotators were not

⁴ <https://elexis.babelscape.com/>

limited to selecting among existing senses; they could also identify missing meanings and extend the inventories with new sense descriptions whenever required. Consequently, the annotation process contributed not only to corpus enrichment but also to the improvement and expansion of the underlying lexical resources.

The multilingual nature of *ELEXIS-WSD* is reflected in the diversity of sense inventories used across languages. The Bulgarian subcorpus relies on the *Dictionary of Bulgarian*, Danish on *DanNet*, English on the *Open English WordNet*, Spanish on *Spanish Wiktionary*, Estonian on the *EKI Combined Dictionary of Estonian*, Hungarian on the *Explanatory Dictionary of the Hungarian Language*, Italian on a combination of *PSC* and *Italian WordNet*, Dutch on the *Open Dutch WordNet*, Portuguese on the *Portuguese Academy Dictionary (DACL)*, Slovene on the *Digital Dictionary Database of Slovene*, and Serbian on *Serbian WordNet*.

Figure 4 presents the distribution of lexical units in the sense inventories across the language subcorpora, grouped by major part-of-speech categories. Nouns and verbs constitute the largest portions of the inventories in all languages, reflecting the central role of these categories in lexical-semantic annotation and word sense disambiguation tasks. Adjectives represent the third most frequent category, while adverbs, proper nouns (for some languages), multiword expressions, and other lexical classes occur in smaller numbers. Considerable variation can be observed in the overall size of the inventories, with Hungarian and Portuguese containing the largest numbers of annotated lexical units, whereas smaller inventories are available for languages such as Estonian and Danish. Despite these differences, the overall distributional patterns remain similar across languages, indicating a comparable semantic organization of the inventories and facilitating cross-lingual lexical-semantic analysis within the *ELEXIS-WSD* framework.

Figure 5 (left) shows the proportion of multiword expressions (MWEs) among all lexical units in the sense inventories used for the *ELEXIS-WSD* corpus. The results reveal substantial variation across languages, reflecting differences in inventory design, and the treatment of multiword lexical units. Serbian exhibits the highest proportion of MWEs (17.5%), followed by Slovene (9.1%) and Danish (8.4%), indicating a strong emphasis on representing phraseological and multiword lexical phenomena within these resources. In contrast, the cur-

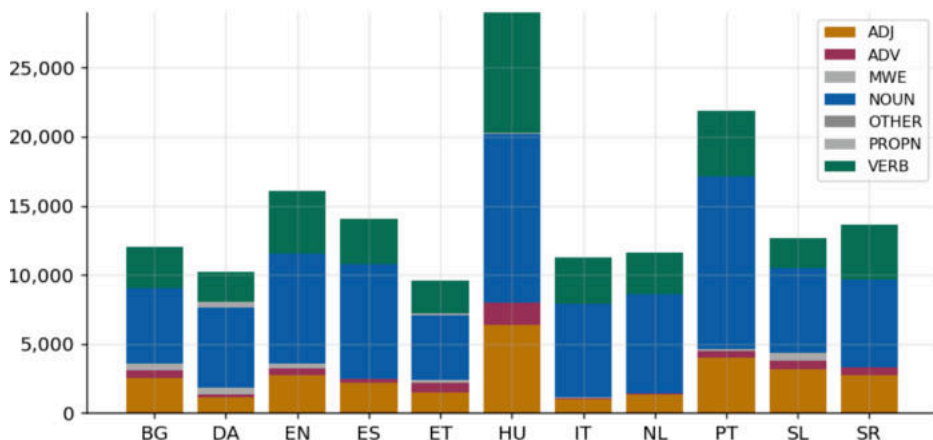


Figure 4: Sense inventory in ELEXIS 2.0.

rent sense inventories for Hungarian (0.1%), Dutch (0.8%), Spanish (0.9%), and Italian (1.2%) contain relatively few MWEs. Figure 5 (right) shows that the majority of MWEs consist of two words, although expressions containing up to eight words are also attested in the inventories.

The observed differences are important for multilingual word-sense annotation, as MWEs often exhibit idiosyncratic meanings that cannot be derived compositionally from their constituent words. Inventories containing a larger proportion of MWEs provide broader coverage of phraseological phenomena and may therefore support more accurate semantic annotation and disambiguation. At the same time, the variability across languages highlights one of the challenges in creating harmonized multilingual lexical resources, since the notion of what constitutes a lexicalized multiword expression is not always consistent across dictionaries and lexical databases. These findings further emphasize the importance of dedicated MWE annotation efforts, such as those carried out within UniDive and PARSEME, for improving the comparability and semantic richness of multilingual lexical resources.

Beyond their role in annotation, sense inventories constitute a valuable research resource in their own right. They provide structured representations of lexical meaning, support the evaluation of word-sense disambiguation systems, facilitate multilingual lexical-semantic studies, and offer high-quality semantic knowledge that can be integrated into modern NLP applications, including knowledge graphs,

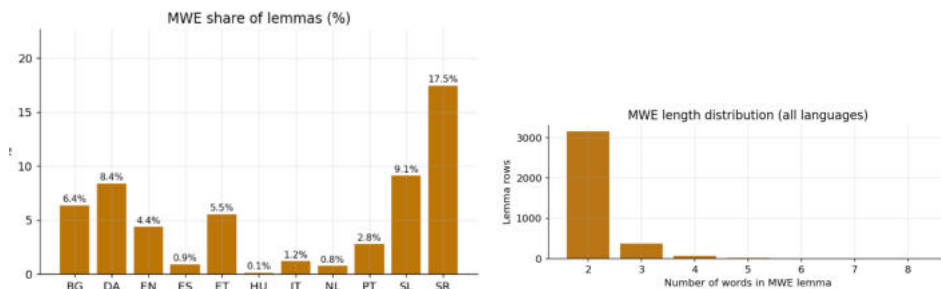


Figure 5: Sense inventory in ELEXIS 2.0.

semantic search systems, and large language models. The continued development and enrichment of these inventories therefore represent an essential prerequisite for advancing multilingual lexical-semantic research and language technology.

Figure 6 presents the distribution of polysemy across the sense inventories used for individual languages in ELEXIS-WSD. In all inventories, the majority of lemmas are associated with only one or two senses, while the number of lemmas decreases rapidly as the number of senses increases. This long-tail distribution is characteristic of lexical resources and reflects the fact that most lexical units have a limited number of meanings, whereas a relatively small set of highly frequent and semantically productive words exhibit extensive polysemy. Despite differences in inventory size, lexicographic tradition, and source resources, the overall shape of the distributions is remarkably similar across languages, indicating comparable approaches to sense granularity. Some inventories, such as those for Hungarian and Portuguese, contain a larger proportion of highly polysemous lemmas, whereas Serbian, Danish, and Dutch show a steeper decline after the first few senses. These differences likely reflect characteristics of the underlying lexical resources rather than inherent linguistic properties.

Table 1 summarizes the length of sense definitions across the inventories used in ELEXIS-WSD. Considerable variation can be observed between languages, reflecting differences in lexicographic traditions and resource design. Portuguese, Spanish, and Hungarian exhibit the longest definitions on average, with mean lengths above ten words, whereas Slovene definitions are substantially shorter, averaging fewer than three words. This is expected, as the current Slovene sense inventory consists of sense indicators (short phrases that only serve to help discriminate between different senses, not

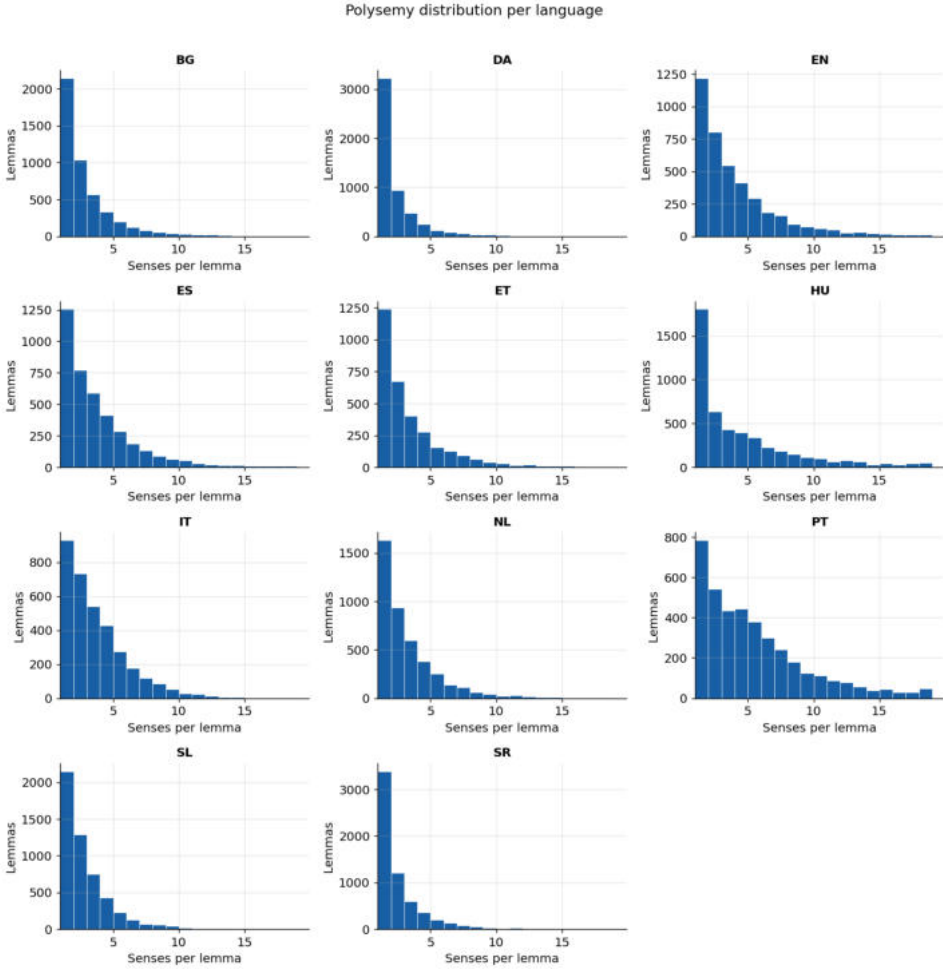


Figure 6: Polisemy in ELEXIS 2.0.

necessarily fully explain them). Most inventories have median definition lengths between six and ten words, suggesting a preference for concise but informative sense descriptions. The standard deviation values indicate varying levels of definition complexity, with Portuguese and Spanish showing the greatest variability and containing some exceptionally long definitions. Portuguese also has the highest maximum definition length (407 words), while Spanish is the only inventory containing a substantial number of empty definitions (260), which may require additional curation before large-scale semantic annotation. Overall, the results highlight the heterogeneity of lexical resources used within ELEXIS-WSD and illustrate one of the chal-

lenges of harmonizing multilingual sense inventories for cross-lingual word-sense annotation.

Table 1: Definition length statistics in the sense inventories used for ELEXIS-WSD.

Lang	Mean	Median	Std	Min	Max	Empty
BG	7.7	7.0	5.3	1	69	0
DA	7.6	8.0	3.1	0	36	2
EN	8.8	8.0	5.3	1	61	0
ES	10.9	9.0	8.5	0	164	260
ET	6.8	6.0	3.8	1	29	0
HU	10.8	10.0	6.9	1	152	0
IT	9.2	8.0	6.6	1	52	0
NL	6.9	5.0	6.3	1	93	0
PT	11.3	10.0	9.5	1	407	0
SL	2.7	2.0	1.2	1	12	0
SR	7.7	7.0	4.6	1	67	0

5. Current State of ELEXIS-WSD 2.0

Figure 7 summarizes the current status of annotation layers across all language subcorpora included in *ELEXIS-WSD* 2.0. The figure provides an overview of both the original language subcorpora inherited from previous versions of the corpus and the newly introduced languages added within the UniDive project. Green cells indicate annotation layers that have been completed, manually verified, and incorporated into version 2.0, whereas orange cells denote annotation activities that are currently in progress and are expected to be finalized by the end of the UniDive project. These additional annotations will be incorporated into future releases of *ELEXIS-WSD*.

The original version of the corpus (version 1.1) contained subcorpora for ten languages and included sentence-level alignment, tokenization, morphosyntactic annotation, and word-sense annotation. Within UniDive, the resource has undergone substantial expansion through the addition of eight new language subcorpora: Greek, Croatian, Georgian, Macedonian, Polish, Romanian, Serbian, and Ukrainian. As a result, the corpus now covers approximately 38,000 sentences and 620,000 tokens, representing an increase of about 80% compared to the pre-UniDive version, which contained approximately 20,000 sentences and 345,000 tokens.

Lang	First in version	Translation	Tokenization	POS & Lemmatization	MWE	MWE by types	NEs	WSD
BG	1							
DA	1							
EN	1							
ES	1							
ET	1							
HU	1							
IT	1							
NL	1							
PT	1							
SL	1							
EL	2							
HR	2							
KA	2							
MK	2							
PL	2							
RO	2							
SR	2							
UK	2							

Figure 7: Current State of *ELEXIS-WSD* version 2.0.

The figure also highlights the gradual and multi-layered nature of the corpus extension process. Linguistic annotation layers, including translation validation, tokenization, lemmatization, morphosyntactic annotation, and dependency parsing, have been completed for most languages. More advanced semantic layers, such as multiword expression annotation, named entity annotation and linking, and word-sense annotation, remain at different stages of completion depending on the availability of linguistic resources and expert annotators. In particular, Serbian currently represents the most advanced newly added subcorpus, including completed named entity annotation and linking as well as an initial word-sense annotation layer. Other languages are expected to reach comparable levels of annotation in future project phases.

Overall, the current state of *ELEXIS-WSD* 2.0 demonstrates both the scalability of the lexicon-corpus interface approach and the feasibility of creating richly annotated multilingual resources through coordinated international collaboration.

6. Conclusion

This paper presented *ELEXIS-WSD 2.0*, a substantially extended version of the *ELEXIS-WSD Parallel Sense-Annotated Corpus* (Čibej et al., 2026), developed within Task 2.2 of UniDive Working Group 2. The new release significantly increases both the linguistics

tic coverage and annotation depth of the resource through the addition of eight new language subcorpora and several new annotation layers. As a result, the corpus has grown to approximately 38,000 sentences and 620,000 tokens, making it one of the largest multilingual sense-annotated resources currently available for lexical-semantic research. The corpus is openly available under the CC BY-SA 4.0 license through the CLARIN.SI repository,⁵ ensuring long-term accessibility and reuse by the research community.

Beyond the quantitative expansion, the project has demonstrated the feasibility of combining linguistic and semantic annotation layers within a unified lexicon-corpus interface. In addition to tokenization, lemmatization, morphosyntactic annotation, and parsing, the corpus now incorporates multiword expression annotation, named entity annotation and linking, and, for Serbian, a fully operational workflow for word-sense annotation. These extensions substantially increase the value of the resource for lexical semantics, word-sense disambiguation, multilingual NLP, digital lexicography, and cross-lingual studies.

Work on the corpus will continue throughout the remaining period of UniDive and will provide an important foundation for future activities within the forthcoming *ELEXAI* project (*European Lexicographic Infrastructure for Artificial Intelligence*, 2026–2029). Particular attention will be devoted to completing and harmonizing the remaining annotation layers across languages, with a special focus on word-sense annotation and the development of high-quality sense inventories. One of the major findings of the project is that the availability of machine-readable, openly accessible lexical-semantic resources remains a significant bottleneck for many languages. Several participating languages either lack comprehensive sense inventories altogether or rely on resources that are incomplete, outdated, or not readily usable for corpus annotation.

Future work will therefore focus on both resource creation and methodological innovation. In particular, we plan to investigate the use of large language models for the semi-automatic generation of missing senses and definitions, the enrichment of existing sense inventories, and the automatic linking of senses across lexical resources and languages. Additional research directions include supersense annotation, cross-lingual sense alignment, knowledge graph integration,

⁵ ELEXIS-WSD 2.0 at CLARIN.SI: <http://hdl.handle.net/11356/2101>

and the development of AI-assisted lexicographic workflows. These efforts will contribute not only to the further expansion of *ELEXIS-WSD*, but also to the broader goal of building interoperable lexical-semantic infrastructures capable of supporting next-generation multilingual language technologies.

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, the European Union’s Horizon Europe research and innovation programme under grant agreement No 101293013, project ELEXAI: The European Lexicographic Infrastructure for Artificial Intelligence.

Language Resources

Čibej, Jaka and Krek, Simon and Tiberius, Carole and Martelli, Federico and Navigli, Roberto and Kallas, Jelena and Gantar, Polona and Koeva, Svetla and Nimb, Sanni and Sandford Pedersen, Bollette and Olsen, Sussi and Langemets, Margit and Koppel, Kristina and Üksik, Tiit and Dobrovoljc, Kaja and Ureña-Ruiz, Rafael and Sancho-Sánchez, José-Luis and Lipp, Veronika and Váradi, Tamás and Györffy, András and Simon, László and Quochi, Valeria and Monachini, Monica and Frontini, Francesca and Tempelaars, Rob and Costa, Rute and Salgado, Ana and Munda, Tina and Kosem, Iztok and Roblek, Rebeka and Kamenšek, Urška and Zaranšek, Petra and Zgaga, Karolina and Ponikvar, Primož and Terčon, Luka and Jensen, Jonas and Flörke, Ida and Lorentzen, Henrik and Troelsgård, Thomas and Blagoeva, Diana and Hristov, Dimitar and Kolkovska, Sia and Muischnek, Kadri and Saul, Kertu and Jõgi, Karoliina and Bon, Mija and Stanković, Ranka and Krstev,

Cvetana and Marković, Aleksandra and Ikonić Nešić, Milica and Giouli, Voula and Papanikolaou, Eri and Lobzhanidze, Irina and Barbu Mititelu, Verginica and Popa, Simina and Cristiana, Lea and Catalin, Mihaila and Irimia, Elena and Ostroški Anić, Ana and Runjaić, Siniša and Sviben, Robert and Pavić, Martina and Filipović Petrović, Ivana and Alberski, Bartłomiej and Cvetkoski, Vladimir and Kanishcheva, Olha and Makhachashvili, Rusudan. 2026. *Parallel sense-annotated corpus ELEXIS-WSD 2.0*. Slovenian language resource repository CLARIN.SI.

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 Györffy, 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 Ped-

- ersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Györffy, 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, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Pais, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Diaz Hernandez, Victoria Beatrix Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, Chaya Liebeskind, Irina Lobzhanidze, Aleksandra Markovic, Gunta Nešpore-Bērzkalne, Adriana Silvina Pagano, Mehrnoush Shamsfard, Ranka Stanković, Vahide Tajalli, Carole Tiberius, and Aakanksha Padhye. 2026. [PARSEME Corpus 2.0 and Shared Task on Automatic Identification of Multiword Expressions](#). In *Proceedings of the 4th UniDive Workshop*, Bucharest, Romania.
- Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanатов, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, 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.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. [WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361,

Online. Association for Computational Linguistics.

- Ranka Stankovic, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. [Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3954–3962, Marseille, France. European Language Resources Association.
- 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*.
- Ranka Stanković, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. [Parallel Bidirectionally Pretrained Taggers as Feature Generators](#). *Applied Sciences*, 12(10).
- 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.
- Jaka Čibej, Simon Krek, Carole Tiberius, Irina Lobzhanidze, Verginica Barbu Mititelu, Jelena Kallas, Kertu Saul, Kadri Muischnek, Karoliina Jõgi, Ranka Stanković, Cvetana Krstev, Aleksandra Markovic, Ana Ostroški Anić, Bartłomiej Alberski, Vladimir Cvetkoski, Voula Giouli, Rusudan Makhachashvili, and Olha Kanishcheva. 2026. [Extension of the ELEXIS-WSD Parallel Sense-](#)

Annotated Corpus Within UniDive: New Languages and Layers.
In *Proceedings of the 4th UniDive Workshop*, Bucharest, Romania.

ELEXIS-WSD 2.0: Proširenje interfejsa između leksikona i korpusa novim jezicima i slojevima anotacije

Jaka Čibej, Simon Krek, Carole Tiberius, Irina Lobzhanidze, Verginica Barbu Mititelu, Jelena Kallas, Kertu Saul, Kadri Muischnek, Karoliina Jõgi, Ranka Stanković, Cvetana Krstev, Aleksandra Marković, Milica Ikonić Nešić, Ana Ostroški Anić, Bartłomiej Alberski, Vladimir Cvetkoski, Voula Giouli, Rusudan Makhachashvili, Olha Kanishcheva

Sažetak

Paralelni korpus sa anotiranim značenjima ELEXIS-WSD predstavlja višejezički resurs namenjen podršci zadacima razrešavanja značenja reči, leksičko-semantičkim istraživanjima i međujezičkim analizama, povezivanjem korpusnih primera sa leksičkim resursima putem ručne anotacije značenja. U okviru COST akcije UniDive, Radna grupa 2 (WG2) bila je posvećena proširenju i unapređenju ovog resursa, što je rezultovalo objavljivanjem verzije ELEXIS-WSD 2.0 u aprilu 2026. godine. Nova verzija značajno proširuje korpus uvođenjem osam novih jezika: rumunskog, grčkog, ukrajinskog, poljskog, makedonskog, gruzijskog, hrvatskog i srpskog, dok su postojeći potkorpusi za slovenački, estonski i holandski dodatno obogaćeni novim slojevima anotacije.

Proces proširenja obuhvatio je više anotacionih slojeva. Novi jezički potkorpusi nastali su automatskim prevodenjem 2.024 izvorne engleske rečenice, nakon čega su usledile ručna korekcija i validacija prevoda. Tekstovi su zatim tokenizovani, lematizovani, morfosintaksički anotirani i sintaksički parsirani, uz ručne ispravke gde su raspoloživi resursi to omogućavali. Dodatni semantički slojevi uključuju anotaciju višechlanih izraza prema smernicama PARSEME 2.0 i anotaciju imenovanih entiteta povezanih sa bazom Wikidata. Srpski tim dao je značajan doprinos kroz anotaciju i povezivanje imenovanih entiteta, kao i izradom prvog potkorpusa sa anotiranim značenjima primenom poluautomatskog postupka anotacije.

Kao rezultat, ELEXIS-WSD 2.0 sada obuhvata približno 38.000 rečenica i 620.000 tokena, što predstavlja povećanje od oko 80% u odnosu na verziju korpusa pre početka projekta UniDive. Korpus je javno dostupan putem repozitorijuma CLARIN.SI i predstavlja značajnu infrastrukturu za višejezička leksičko-semantička istraživanja. Budući rad biće usmeren na završetak preostalih anotacionih slojeva, harmonizaciju anotacija između

jezika, proširenje inventara značenja i istraživanje mogućnosti primene velikih jezičkih modela za automatsko generisanje i povezivanje značenja.

Ključne reči: razrešavanje značenja reči, paralelni korpus, semantička anotacija, imenovani entiteti, višečlani izrazi, leksički resursi, ELEXIS-WSD

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

