SrELEXIS-WSD: Hybrid Semi-Automated WSD for Serbian with Large Language Models—Results and Challenges

Ranka Stanković[1,2], Cvetana Krstev[1], Saša Petalinkar[1,2], Milica Ikonić Nešić[3], Jelena Bogdanović[3], Milica Dinić Marinković[3], Aleksandra Marković[4], Marina Bagi[4], Marijana Đukić[4]

[1]Jerteh-Language Resources and Technologies Society, Belgrade, Serbia
[2]University of Belgrade, Faculty of Mining and Geology, Belgrade, Serbia
[3]University of Belgrade, Faculty of Philology, Belgrade, Serbia
[4]Institute for the Serbian Language SASA, Belgrade, Serbia

ranka@rgf.rs; {cvetana,sasa.petalinkar,milica.ikonic}@jerteh.rs
{bogdanoviccjelena,dinicmidata,malexa39,marina.bagi17,marijanadjukic01}@gmail.com

Relevant UniDive working groups: WG2

1 Introduction

The Parallel Sense-Annotated Corpus ELEXIS-WSD (Martelli et al., 2023) has been extended within the UniDive COST Action with new layers and new languages, including subcorpora for South Slavic languages (Čibej et al., 2025). The extension of the ELEXIS-WSD corpus, ELEXIS-WSD-sr was developed by translating 2,024 English WikiMatrix sentences into Serbian, followed by extensive manual correction by multiple native speakers to ensure high quality. The data were automatically tokenized, lemmatized, and POS-tagged, then manually validated by at least three annotators. The resulting corpus contains 31,058 tokens, with an average sentence length of 15.34 tokens (Krstev et al., 2024a). Previous activities were also focused on Multiword Expressions (MWEs) and Named Entities layers, as well as on sense inventory establishment (Krstev et al., 2025). MWE annotations were carefully reviewed following PARSEME guidelines.[1]

2 Methodology for WSD

Word sense disambiguation (WSD) is the task of assigning the most appropriate meaning (sense) to a word (or MWE) in a given context from a predefined sense inventory. For the final stage in constructing ELEXIS-WSD-sr, namely WSD, we applied a semi-automated methodology. MWEs were resolved first, after that content words (nouns, verbs, adjectives, adverbs) were processed by matching their lemma and POS against a curated sense inventory based on Serbian WordNet (Krstev et al., 2024b, 2025) to retrieve candidate senses. Each instance is embedded in a structured prompt A containing the context, lemma, and candidate senses, with the target word highlighted. A chosen LLM performs zero-shot disambiguation, returning a constrained JSON output with the predicted sense and a brief explanation. The response is parsed, and both the sense label and explanation are reintegrated into the sentence annotation. The process is applied across the corpus, producing a sense-annotated dataset aligned with the inventory and ready for manual evaluation in INCEpTION. [2]

The prompt is tailored for Serbian, using glosses from the sense inventory. Multilingual prompting was avoided due to instability in low-resource cross-lingual settings.

3 Corpus preparation

The ELEXIS-WSD-sr corpus includes 13,351 sense-assigned single-word instances and 1,242 MWEs, distributed across four UPOS categories: nouns (single 6,513/mwe 727), verbs (s. 2,491/mwe 428), adjectives (s. 3,358/mwe 5), and adverbs (s. 988/mwe 82). An evaluation of the automatic methods used for MWE identification and classification is presented in (Krstev et al., 2025).

4 Sense Inventory

The Serbian sense inventory (WSD-sr-sens-inv) was built on SrpWN (Krstev et al., 2004; Stanković et al., 2018) and extended to support large-scale WSD by aligning it with the English ELEXIS-WSD inventory based on Princeton WordNet (PWN). Missing synsets were identified, automatically translated, and manually corrected and refined, resulting in the addition of over 2,000 new

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

[2] http://inception.jerteh.rs/

synsets and the expansion of SrpWN to nearly 30,000 synsets.

Further enrichment included extracting content words from the corpus, adding missing senses via translation, and adding dictionary definitions and LLM-generated entries, all of which were subsequently validated manually. For instance, a sense missing in SrpWN (and PWN) was *mis* 'Miss', for which the definition from a Serbian dictionary was added: *devojka koja je izabrana kao najlepša na nekom konkursu lepote, pobednica na takvom konkursu* 'a girl who was chosen as the most beautiful in a beauty contest, the winner of such a contest'. Another example is *čet bot* 'chatbot' for which an LLM provided a definition: *Kompjuterski program dizajniran za simulaciju razgovora sa ljudima, obično putem teksta.* 'A computer program designed to simulate conversations with humans, usually through text.' The inventory currently covers most corpus-relevant senses and continues to be extended.

The Serbian sense inventory was converted into an RDF-based knowledge base (KB) for the INCEpTION platform, enabling structured, explainable WSD and direct linking between senses and corpus instances. Lemmas were modeled as classes linked to POS categories, while senses were represented as instances enriched with definitions, synonyms, and metadata, supporting a synset-like structure.

The KB facilitates transparency, manual validation, and iterative refinement by allowing inspection and editing of senses alongside real corpus examples. It thus functions as a dynamic resource that supports both sense inventory management and continuous evaluation of WSD annotations.

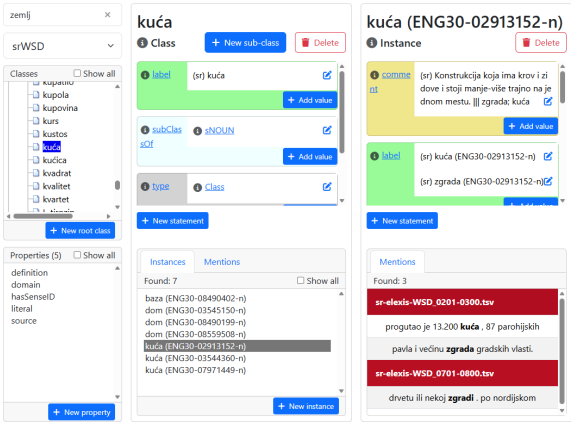

Figure 1: An illustration of the KB in INCEpTION

Figure 1 shows the INCEpTION Knowledge Base interface for managing the Serbian WSD inventory. On the left, a hierarchy of lemma classes is displayed, with *kuća* 'house' selected as a class linked to the NOUN category. The central panel presents its properties, including the label in Serbian and class relations. On the right, a specific sense instance (*kuća*, ENG30-02913152-n) is shown with its definition, labels, and associated synonyms. The lower panel lists related sense instances and corpus mentions, illustrating how lexical senses are connected to annotated examples.

5 Sense Annotation

INCEpTION integrates lexical knowledge with corpus evidence, allowing annotators to manage senses and validate assignments directly in context. The annotation framework integrates multiple layers (POS, lemmas, Named Entity Recognition (NER), Named Entity Linking (NEL), MWE, WSD), with LLM-based sense assignments supported by explicit natural-language explanations. Each sense is linked to a URI and justified in context, ensuring transparency and traceability, while named entities are linked to Wikidata.

After the initial manual correction, the updated inventory (WSD-sr-sens-inv-v2) was expanded to include previously missing senses identified in the corpus and in the error analysis (where the system was unable to find a suitable sense), thereby improving subsequent automatic annotation. LLM-generated explanations supported manual revision by highlighting mismatches and gaps in definitions. In addition, 213 sense definitions were refined for clarity and precision. The update mainly increased sense coverage and granularity for existing lemmas, with minimal changes to the overall lexical structure.

6 Results and Evaluation

The study evaluated multiple LLMs (GPT-3.5, GPT-4.1, GPT-5.1, Gemini 2.5 Flash, Gemini 2.5 Pro, Llama 4, and Mistral Small 3.2) for Serbian WSD under a unified prompting setup, alongside embedding-based baselines (TeslaXLM, MiniLM, and multilingual E5) and a knowledge-based Lesk baseline. The experiments were conducted in two rounds: first with the initial inventory, followed by manual evaluation and expansion, which improved coverage and enabled a second round of annotation with the expanded inventory.

Evaluation on a manually curated gold dataset split into a 120-sentence development/analysis set (772 annotated senses) and a 300-sentence held-out test set (1,921 annotated senses) showed that LLM performance improves with a richer and improved sense inventory, with GPT-4.1 emerging as the strongest model. All of this led to inventory expansion, refinement of the senses, and improved handling of challenging cases, such as auxiliary-like verbs and MWEs.

Model performance was analyzed by comparing correctly matched, mismatched, and missing senses. Results show that LLMs consistently outperform baseline models across both rounds, and that improvements in the sense inventory substantially increase accuracy for all models. In second round, GPT-4.1 achieved the best overall performance on the held-out test set (accuracy = 0.927; 1,780 matched / 130 mismatched / 11 unmatched senses), outperforming GPT-5.1, while all models demonstrated clear gains in the second round due to the enhanced sense inventory.

The linked data version of the corpus and sense inventory is available at GraphDB SPARQL endpoint[3].

7 Conclusion and future work

The study shows that LLMs effectively support WSD in Serbian and benefit from improved sense inventories. However, full automation remains unattainable, as complex and culturally specific cases still require expert validation. Errors often stem from limitations in the sense inventory rather than the models, highlighting the need for better coverage and clearer definitions. Explanation-based prompting improves performance and transparency and helps identify inventory gaps. Overall, the workflow enables both WSD evaluation and iterative lexicon enhancement and is transferable to other low-resource languages.

Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia, #GRANT 7276, Text Embeddings-Serbian Language Applications – TESLA, the Ministry of Science of the Republic of Serbia (451-03-33/2026-03/200174), and COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive).

[3] http://graphdb.jerteh.rs/sparql?repositoryId=elexis-wsd-nif and Turtle files (*.ttl*) with descriptions are available at https://llod.jerteh.rs/UNIDIVE/NIF/

References

Jaka Čibej, Ranka Stanković, Ana Ostroški Anić, Simon Krek, and Carole Tiberius. 2025. The ELEXIS-WSD Parallel Sense-Annotated Corpus and South Slavic Languages: Subcorpora for Croatian, Serbian, and Slovene. In *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig: Thematic Collection of Papers*, pages 45–59. Beograd: University of Belgrade—Faculty of Philology.

Cvetana Krstev, Ranka Stanković, and Aleksandra Marković. 2024a. Annotation of MWEs and NEs in the Serbian extension of ELEXIS-WSD: comparisons, solutions and open questions. In *2nd UniDive Workshop in Naples*. COST.

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*.

Cvetana Krstev, Ranka Stanković, Aleksandra M Marković, and Teodora Mihajlov. 2024b. Towards the Semantic Annotation of SR-ELEXIS Corpus: Insights into Multiword Expressions and Named Entities. In *Proceedings of (MWE-UD)@ LREC-COLING 2024*, pages 106–114.

Cvetana Krstev, Duško Vitas, Gordana Pavlović-Lažetić, and Ivan Obradović. 2004. Using Textual and Lexical Resources in Developing Serbian WordNet. *Romanian Journal of Information Science and Technology*, 7(1-2):147–161.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, and ... 2023. Parallel Sense-Annotated Corpus ELEXIS-WSD 1.1.

Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. Resource-based WordNet Augmentation and Enrichment. In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114, Sofia, Bulgaria.

A Appendix: Prompt design

Prompt in English (for illustrative purposes only):
System Message:

You are an expert in lexicon and semantics. Based on the sentence context and the list of valid senses for the target expression, your task is to identify the sense that is most precisely used in the given context.

The response must be provided in a strictly defined JSON format:

{{

"sense_id": "<one of the provided IDs or 'NEW_SENSE'>",

"explanation": "<a brief and clear justification in one or two sentences explaining why this sense applies. If 'NEW_SENSE' is used, explain why none of the provided senses are appropriate.>"

}}

Do not add any text outside the JSON structure. Use 'NEW_SENSE' only if none of the senses is even approximately correct in context. The sense_id must exactly match one of the provided IDs or use 'NEW_SENSE'; you must never invent other IDs.

User Message:

Context: Do kraja godine preostalo je 357 dana (358 u **prestupnoj godini**) . (eng. *There are 357 days remaining until the end of the year (358 in leap years).*

Target expression: prestupna godina (eng. *leap year*)

Possible senses:

1. ID: ENG30-15202230-n — U Gregorijanskom kalendaru svaka godina deljiva sa 4, osim godina deljivih sa 100 koje nisu deljive sa 400. (Engl. *in the Gregorian calendar: any year divisible by 4 except centenary years not divisible by 400*)