Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities

Cvetana Krstev, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov

Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities

Cvetana Krstev, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov
Association for Language Resources and Technologies, Univ. of Belgrade, F. of Mining and Geology,
Institute for the Serbian Language SASA, Association for Language Resources and Technologies
cvetana@jerteh.rs, ranka@rgf.bg.ac.rs, malexa39@gmail.com, teodoramihajlov@gmail.com

Abstract

This paper presents the work in progress on ELEXIS-SR corpus, the Serbian addition to the ELEXIS multilingual annotated corpus (Martelli et al., 2023), comprising semantic annotations and word sense repositories. The ELEXIS corpus has parallel annotations in ten European languages, serving as a cross-lingual benchmark for evaluating low and medium-resourced European languages. The focus in this paper is on multiword expressions (MWEs) and named entities (NEs), their recognition in the ELEXIS-SR sentence set, and comparison with annotations in other languages. The first steps in building the Serbian sense inventory are discussed, and some results concerning MWEs and NEs are analysed. Once completed, the ELEXIS-SR corpus will be the first sense annotated corpus using the Serbian WordNet (SrpWN). Finally, ideas to represent MWE lexicon entries as Linguistic Linked-Open Data (LLOD) and connect them with occurrences in the corpus are presented.

Keywords: multiword expression, named entity, word sense disambiguation, sense repository, LLOD

1. Introduction

Even in the current era of neural language models, there is a high demand for high-quality, openly accessible corpora that are annotated with senses, especially for training and evaluating semantically related NLP tasks, like word sense disambiguation (WSD) and natural language understanding (NLU) (Pedersen et al., 2023b). Despite many efforts in the field over the past decades, such corpora are still scarce for many languages with limited resources, including Serbian. This scarcity is caused not only by the lack of freely available sense inventories, which are necessary for these tasks, but also by the complexity and cost of compiling annotations since it requires substantial manpower, preferably from experienced linguists or lexicographers. For Serbian, the availability of curated dictionaries for such use is limited, and not even subsets for particular corpora annotation are available.

The paper is structured as follows: In Section 2 we give an account of related work, and continue by presenting in Section 3 the ELEXIS-WSD dataset, its extension with Serbian data and its basic annotation layers prior to the semantic annotation. Section 4 discusses the annotation of MWEs and NEs in the ELEXIS-WSD as well as in its Serbian extension. The building of the sense inventory for Serbian and the role of MWEs and NEs in it are presented in Section 5. Possible ideas for publishing dictionaries of MWEs as LLOD and associating its entries with corresponding occurrences in the corpus are developed in Section 6. Finally, in Section 7 we conclude and discuss open questions, potential future research, and development.

2. Related work

A semantic concordance is a textual corpus and a lexicon, combined so that every substantive word in the text is linked to its appropriate sense in the lexicon (Miller et al., 1993). The popularity of SemCor (Landes et al., 1998), one of the initial sense-annotated English corpora based on the Princeton WordNet sense inventory (Fellbaum, 1998) inspired the NLP community to build sense-annotated corpora for many languages. Exploiting parallel texts in the creation of multilingual semantically annotated resources produced the MultiSemCor Corpus (Bentivogli and Pianta, 2005). Another important multilingual sense annotated corpus is the Ontonotes (Weischedel et al., 2011), that uses the WordNet for sense annotations of the English part, whereas the Chinese and Arab parts base the sense annotations on various lexical sources.

The semiautomatic approaches to sense annotation were applied to overcome the scarcity of such data sets. The OneSec are sense-annotated corpora for word sense disambiguation in multiple languages and domains (Scarlini et al., 2020) that consist of Wikipedia texts containing between 1.2 and 8.8 M sense annotations of nouns per language.

The FrameNet project (Baker et al., 1998), based on the idea of describing lexical items through semantic frames, produces semantic frames (which contain information about the se-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 106–114. May 25, 2024.

mantic and syntactic valence of words). Target words are mostly nouns, adjectives, and verbs. Every frame and frame element is accompanied by a set of representative sets of manually annotated corpus attestations, and for every frame, the set of relations it enters is presented. Lexical databases based on the FrameNet principles were (or are being) built for several languages. The Salsa project (Burchardt et al., 2009) produced a German lexicon based on the FrameNet semantic frames and annotated a large German newswire corpus.

The English part of Ontonotes was annotated with verbal MWEs (Kato et al., 2018). The main outcome of the COST action PARSEME were unified annotation guidelines, and a corpus of over 5.4 million words and 62 thousand annotated VMWEs in 18 languages (Savary et al., 2018). Development was continued afterward with the inclusion of more languages and the enlargement of corpora for existing languages. The current edition of PARSEME corpus[1] contains 26 languages, including Serbian (Savary et al., 2023). The expansion of MWE annotations to nominal and other MWEs is the task within COST action UNIDIVE[2].

Named Entity Recognition (NER) enables the identification and classification of key information in text. The most frequently annotated classes are persons, locations, and organizations, but for a deeper text understanding identification of events, roles, time, measures, etc. is also necessary. In addition to that, named entity linking (NEL), also known as disambiguation, normalization, or entity resolution, involves aligning a textual mention of a named entity to an appropriate entry in a knowledge base, assigning a unique identity to mentioned entities.

3. The extension of ELEXIS-WSD

ELEXIS-WSD is a parallel sense-annotated corpus in which content words (nouns, adjectives, verbs, and adverbs) have been assigned senses for 10 languages: Bulgarian (BG), Danish (DA), English (EN), Spanish (ES), Estonian (ET), Hungarian (HU), Italian (IT), Dutch (NL), Portuguese (PT), and Slovenian (SL).[3] The list of sense inventories is based on WordNet for DA (Pedersen et al., 2023a), EN, IT, NL, Wiktionary is used for ES, and national digital dictionaries are used for BG, ET, HU, PT, and SL (Federico et al., 2021).

In order to join this task and obtain the Serbian corpus as a part of the future edition of the sense repository being developed within WG2.T2 of the UniDive, the set of sentences from WikiMatrix[4] in EN was translated automatically (Google translation) into SR. We opted for the automatic translation in order to fasten the process, with the full awareness of the need to manually check the translation afterwards. This process was highly demanding, in terms of time and manpower, but it was an unavoidable step for getting high-quality dataset. A few (eight precisely) Serbian native speakers checked the Serbian sentence set thoroughly, in order to avoid literal or incorrect translation, and after that sentences were read carefully once again to resolve different issues: literally or incorrectly translated MWEs, unresolved references in the text (pronouns, e.g., in SR differ for gender, number, and case, and if the pronoun refers to an NP from the previous context, its reference had to be checked to choose the right morphological form); besides, it was necessary to check phonetic transcriptions of names (particularly personal ones), since in SR proper names are not written in the original form (the second reading and issue-resolving was done by two people). The process was time-consuming because of the very nature of the set – sentences are out of context, full of terms from different scientific areas, many of which are MWEs), their content is of encyclopedic sort, and often it was necessary to read the original document in English and/or some other language to understand the meaning and represent it correctly in SR.

After this process, the set was automatically tokenized, lemmatized, and POS-tagged (Stanković et al., 2020; Stanković et al., 2022). The outcomes of all these automatic procedures are being manually corrected. Results show that 2024 sentences in Serbian dataset have 25,478 word forms, content words tagged as: NOUN – 7,198 (diff. 2,413), PROPN – 1,552 (diff. 1,057), ADJ – 3,291 (diff. 1,256), VERB – 3,121 (diff. 913), ADV – 900 (diff. 287).[5] Tasks that remain to be done include the annotation of MWEs and NEs (the first results are presented in the following section), the syntactic annotation, and linking with the sense repository (the first results are presented in Section 5).

4. MWEs and NEs in WSD

In this section, we are focusing on the annotation of MWEs and NEs in the ELEXIS-WSD and in its Serbian extension. As it will be shown, the number of MWEs and NEs annotated in 10 language sentence sets was not even, probably due to different resources used for their annotation.

[1] https://gitlab.com/parseme/

[2] https://unidive.lisn.upsaclay.fr/

[3] https://www.clarin.si/repository/xmlui/handle/11356/1842

[4] https://ai.meta.com/blog/wikimatrix/

[5] This figures are not final since the annotation is presently being double-checked and harmonized with UD.

In order to compare MWEs and NEs occurring in the whole repository, MWEs and NEs in each of 10 languages were automatically translated into SR (as phrases, not word-to-word), and the number of the same translations obtained by translating MWEs from different languages was calculated. MWEs/NEs were automatically translated into Serbian in order to facilitate the comparison with MWEs/NEs retrieved in the Serbian set. Note that the translation of a MWE into SR need not be a MWE. For instance, *prime minister* (EN) → *premijer* (SR).

4.1. MWEs in ELEXIS-WSD

The number of annotated MWEs in the initial WSD repository is presented in Figure 1). The blue columns present the number of unique lemmas in the WSD, while the orange columns present the number of unique senses. This graphic shows that the numbers of MWEs in the WSD repository differ significantly between languages.

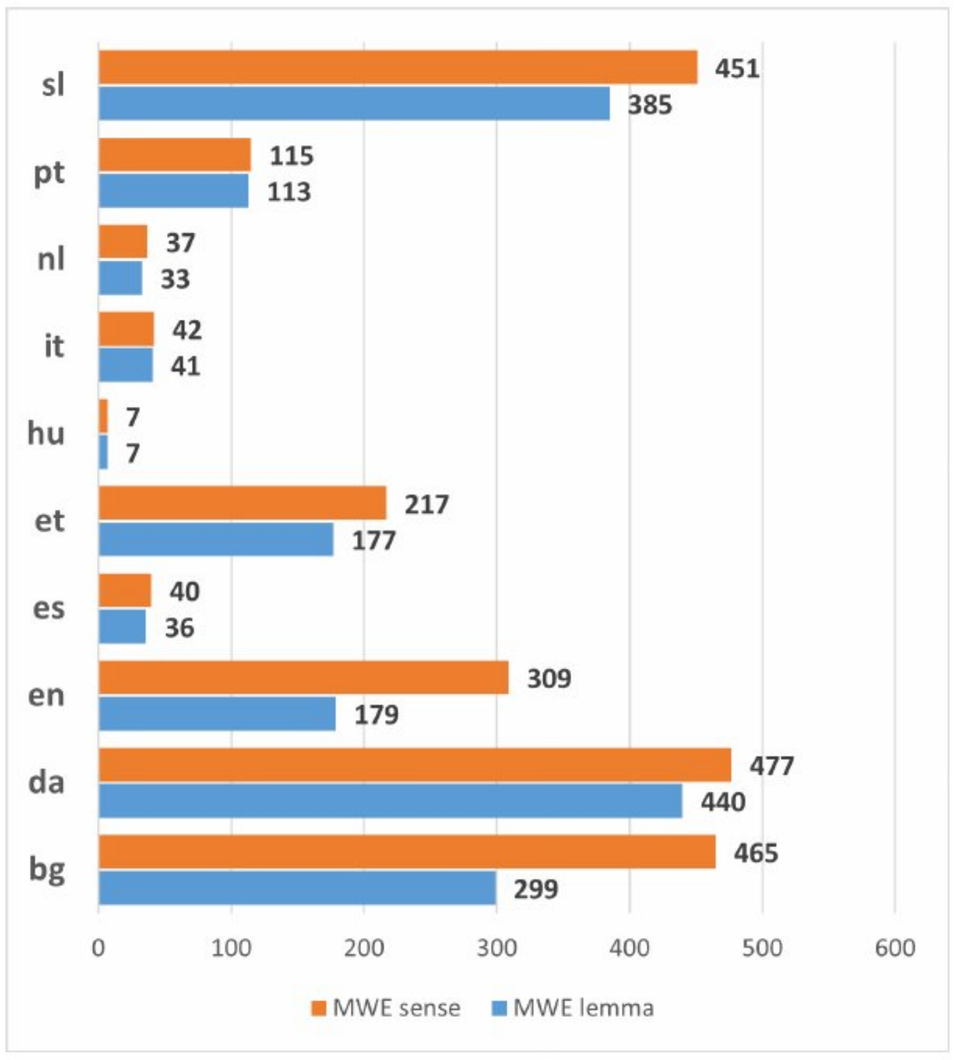

Figure 1: Number of MWEs in the repository – a total of 1,710 MWEs in 10 languages

.

Figure 2 shows that 1,412 different translations were obtained by translating a total of 1,710 MWEs. One international MWE appeared in 6 language sets, *lingua franca*. One of 14 MWEs translated from 4 languages into one SR term was *očekivano trajanje života* (SR): *life expectancy* (EN), *expectativa de vida* (PT), *pričakovana življenjska doba* (SL), *oodatav eluiga* (ET).[6]

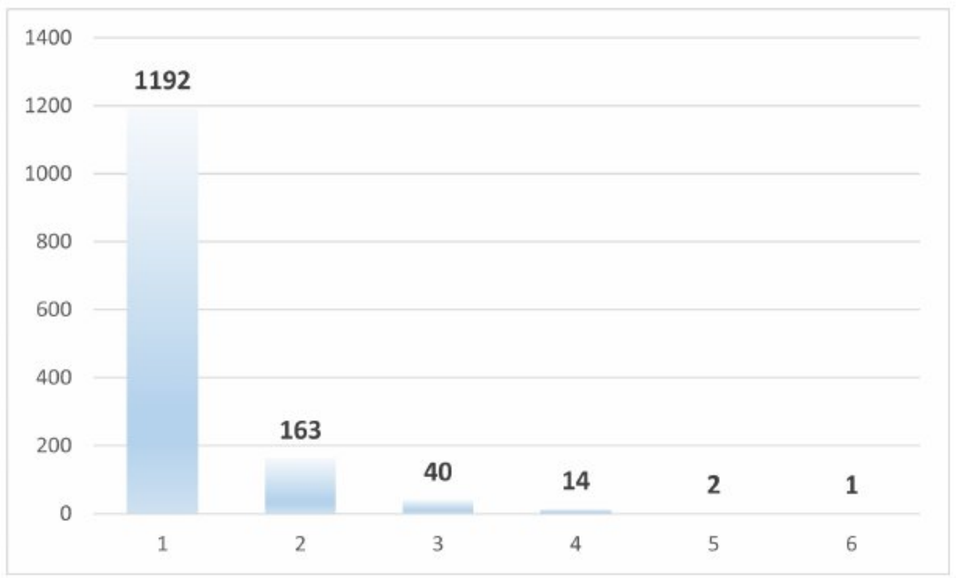

Figure 2: MWEs translations into Serbian obtained by translating from 10 languages – no translation was obtained by translating from more than 6 languages

.

4.2. NEs in ELEXIS-WSD

The number of annotated NEs, without information about specific NE types, is presented in Figure 3 (blue columns present the number of unique lemmas, while orange columns present the number of unique senses).

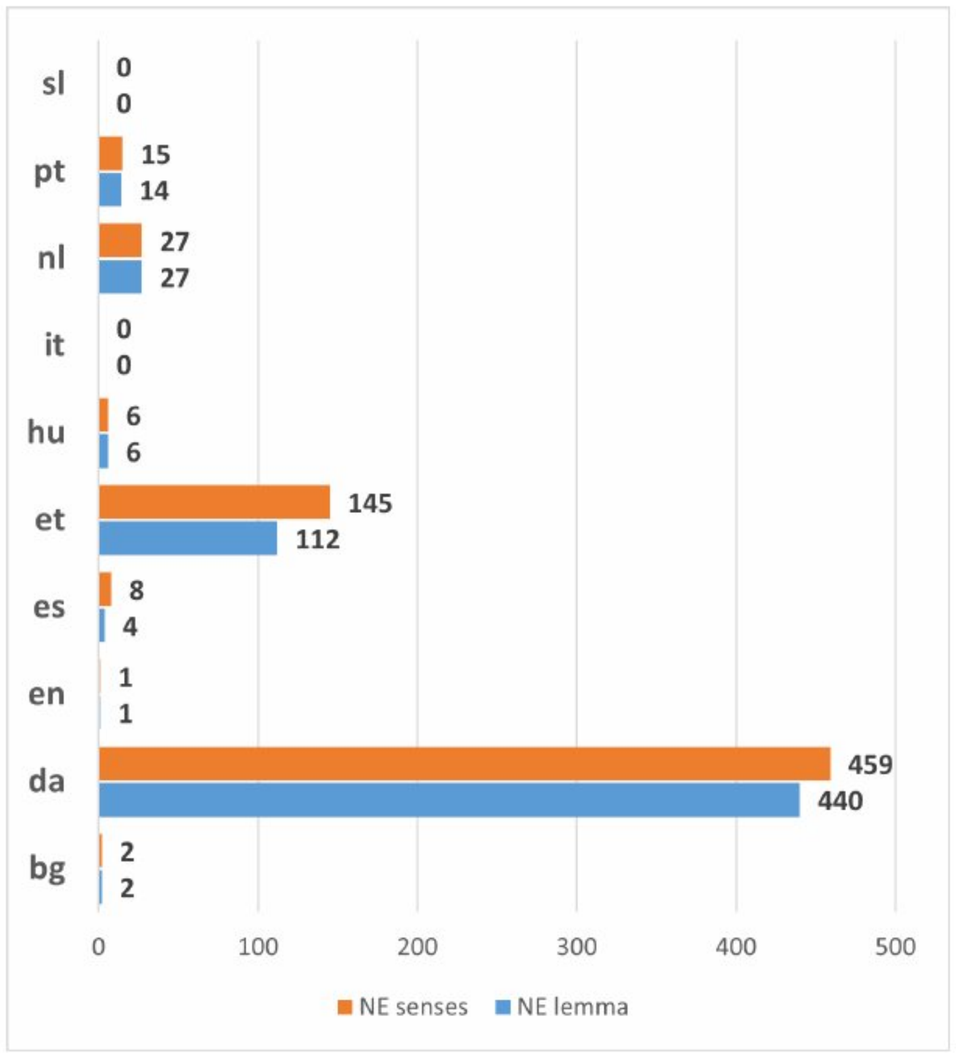

Figure 3: Number of NEs in the repository – a total of 606 NEs in 10 languages

.

Named entities were not systematically annotated in all language datasets (for example, (SL) and (IT) sets have no NE annotated at all, while

[6]The automatic translation was not literal, as demonstrated by the example *srednja škola* (SR) 'lit. middle school' ↔ *high school* (EN); on the other hand it was not always accurate, as demonstrated by *srednja škola* (SR) ↔ *visoka šola* (SL).

some languages have many of them), resulting in 526 translations from a total of 606 NEs. The most frequent NE was *Grčka*, translated from four languages: *Grækenland* (DA), *Grecia* (ES), *Kreeka* (ET), *Grécia* (PT), followed by NEs translated from three languages, one of which is *SAD*: *USA* (EN), *ZDA* (SL), *EUA* (PT). Figure 4 presents the number of translations into Serbian.

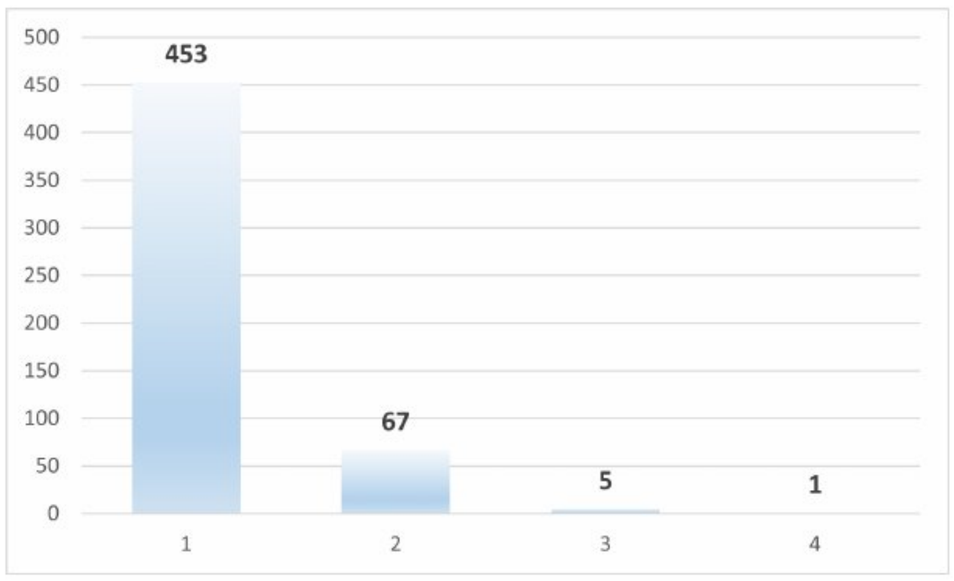

Figure 4: NEs translations into Serbian obtained by translating from 10 languages – no translation was obtained by translating from more than 4 languages

.

4.3. Annotation of MWEs and NEs in the Serbian dataset

The pipeline for preparation and annotation of the Serbian set of 2,024 sentences is presented in Figure 5 – green color boxes and the closed locker symbol represent the finished tasks, pink color boxes and the open locker symbol designate the work in progress, mostly in the evaluation phase, while the pending tasks or tasks in their initial phase are represented by lilac boxes.

The Serbian set of 2,024 sentences was automatically annotated using four different resources and tools:

- The e-dictionary of non-verbal MWEs was used for the annotation of such MWEs. This dictionary was built on the same principles used for building the e-dictionary of simple words for Serbian. The inclusion of MWEs in this dictionary was based on several rather loose criteria: their appearance in some general, terminological or phraseological dictionary of Serbian as well as SrpWN, the frequency of their occurrence in corpora of Serbian, and the intuition of the resource author. The application of this resource to the Serbian sentence set resulted in 529 annotations (339 different) (Krstev et al., 2013). Among them were 351 (249) nominal MWEs, 133 (70) proper nouns, 44 (19) adverbial, and one adjectival.

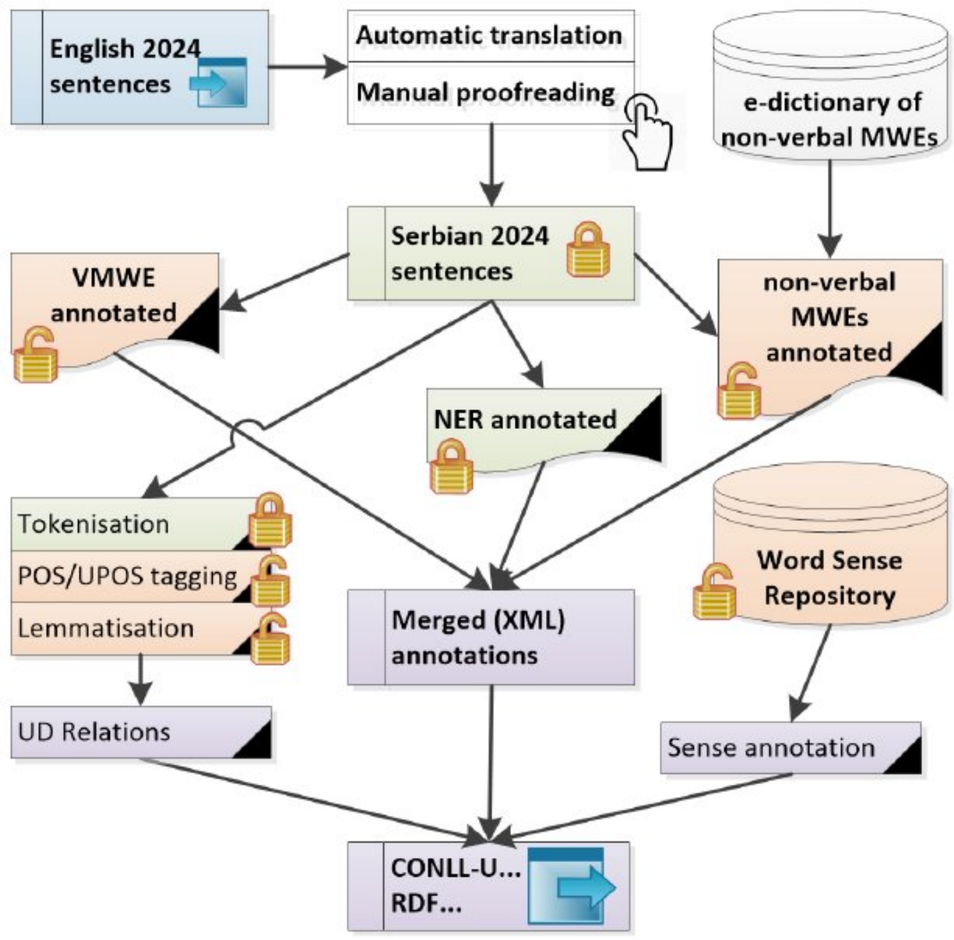

Figure 5: The pipeline for the preparation of the Serbian dataset

.

- A system for the Named Entity Recognition (NER) based on e-dictionaries and rules annotated 2,006 occurrences of NEs (Krstev et al., 2014). Numbers of recognized NEs per class are presented in Table 1. Some multiword named entities, particularly organization (ORG) and geopolitical names (TOP), are recognized both by dictionaries and the NER system.

- A system for the recognition of verbal MWEs based on e-dictionaries, rules, and the repertoire of VMWEs annotated in the Serbian part of the PARSEME Corpus Release 1.3 (Savary et al., 2023) annotated 230 occurrences of VMWEs (98 different), distribution by type: IRV – 174 (62), LVC.full – 35 (21), VID – 13 (10), and LVC.cause – 8 (5).

- A system for the recognition of adjectival and verbal similes is based on a set of more than 600 adjectival and more than 300 verbal similes. It can retrieve different variances of these similes, both in lexica and structure (Krstev et al., 2023). The previous research established that in literary texts an average of 2.2 adjectival similes can be expected per 10,000 words of a text; however, this system in the Serbian sentence set did not retrieve even a single one (Krstev, 2021).

The accuracy of MWE/NE recognition, recall, and precision will be determined during the next

step, when senses will be associated with simple- and multi-word units. Previous evaluations of used systems for the recognition of NEs and MWEs (Krstev et al., 2013; Šandrih et al., 2019) have shown that these systems prioritize precision over recall, which means that in the later stages of processing, through comparison with annotations in datasets for other languages and manual evaluation, new entities will be annotated. It is to be expected that the assignment of senses will reveal some additional MWEs and NEs. This, in turn, will enable the enhancement of used resources and procedures.

Tag	№	Tag	№
PERS	329	TIME	372
TOP	448	AMOUNT	169
ORG	126	MEASURE	62
DEMONYM	244	PERCENT	51
ROLE	175	MONEY	12
EVENT	18	**Total**	2,006

Table 1: Recognized NEs by classes.

4.4. The comparison of MWEs and NEs across languages

Our initial comparison of MWEs and NEs annotated in the WSD repository and in the Serbian sentence set (ELEXIS-SR) was based on their automatic translation to SR, as explained in Subsection 4. This was not ideal, since in several cases the translation was not appropriate: e.g., the SR highly polysemous verb *dovesti* was obtained as a translation equivalent of two VMWEs from two languages, appearing in two unrelated sentences: *dado lugar, 'leed to'* (ES: 700) and *tog med,'take in'* (DA: 148). Once the automatic translation was checked, as actual equivalents of these VMWEs in ELEXIS-SR appeared to be *primiti* (148) ('take in' in ELEXIS-EN), and *dovesti* (700) ('leed to' in ELEXIS-EN), the translated verb itself.

On the other hand, in many cases automatic matches were good: e.g., the SR translation *bruto domaći proizvod* 'gross domestic product' was obtained from MWEs in four languages: *gross domestic product* (EN), *produto interno bruto* (PT), *bruto domači proizvod* (SL), *sisemajandus koguprodukt* (ET), all occurring in the same sentence – 1258. In the corresponding sentence in ELEXIS-SR the translated term *bruto domaći proizvod* was used and annotated as MWE. In all mentioned languages these terms were also annotated as MWEs (in ELEXIS-SL only its part is annotated: *domači proizvod*).

In other cases, the translation was good, it was used in ELEXIS-SR, but it was not annotated in it because it was missing in the used resources. This was the case for *prirodna selekcija*, translated from *natural selection* (EN), *seleção natural* (PT), *naravni izbor* (SL), used in sentence 1560 in ELEXIS-SR, but not annotated in it. This case of missing annotations occurs in other languages as well. E.g., equivalents for *gross domestic product* are MWEs in BG, ES, HU, NL, but yet are not annotated. In IT an acronym was used instead, and in DA a compound.

Having all this in mind, the overall results of the comparison are as follows: out of 653 non-verbal MWEs occurrences (384 lemmas) annotated in ELEXIS-SR, 116 MWE lemmas occurred in at least one language set in WSD; out of 228 VMWE occurrences (99 lemmas) annotated in ELEXIS-SR, 11 lemmas occurred in at least one language set; only 93 NEs annotated in ELEXIS-SR were annotated as MWE or PROPN in WSD (maybe due to the poor lemmatization, automatic translation and linking of proper names).

5. Sense repository

Since there is no freely available digital descriptive dictionary of the Serbian language, the Serbian sense repository will be based on the Serbian WordNet SrpWN (Stanković et al., 2018). ELEXIS sense repository for English, which is also based on the Princeton WordNet (PWN), has 16,106 entries, each assigned with its internal identifier. Since the WordNet interlingual index is not available in the ELEXIS-EN sense repository, we aligned PWN synsets with the ELEXIS-EN sense repository entries by comparing their definitions. This process yielded 13,703 matches.

Subsequently, synsets from this subset were aligned with the Serbian WN containing 25,322 synsets, which revealed that there were 5,997 matches. Finally, the subset missing from the list of 13,703 synsets was compared with sentence annotations in ELEXIS-EN, which revealed that the "urgent" first step is to fill the gap with 2,130 synsets. After the automatic translation of this list of synonyms and their definitions from the PWN using Google API and OpenAI services, the obtained list of Serbian candidates was expanded using several other lexical resources compiled in previous research. Postediting the list of synonym set candidates and their definitions is an ongoing activity.

The further analysis showed that from 437 MWEs annotated in ELEXIS-SR (see Subsection 4) – 339 non-verbal and 98 verbal – 171 (39%) were found in the Serbian WordNet. Moreover, some of them occur in 2 or more synsets. Table 2 gives the total number of senses and the number of lemmas (literals) per MWE type. For instance, *komunikacioni sistem* 'communication system' can refer to a '(def.) system for communicating' or to

Group	Type	Senses	Lemma
MWE	NOUN	100	94
MWE	PNOUN	35	32
VMWE	IRV	80	42
VMWE	LVCfull	1	1
VMWE	VID	2	2

Table 2: Annotated MWEs in ELEXIS-SR retrieved in SrpWN per type.

'(def.) a facility consisting of the physical plants and equipment for disseminating information.' The VMWE with the highest number of different meanings is the reflexive verb *pojaviti se*: 'to appear' (to come in sight), 'to come up' (of celestial bodies), 'to come out' (be issued), 'to originate' (come into existence), 'to arise' (result or issue). A proper noun having two senses is *Novi Zeland* 'New Zealand', referring to a country and an island (same as in the PWN). Besides reflexive verbs, one light verb construction is recorded in the subset of SrpWN potentially interesting for our research – *dati ostavku* 'give resignation' – and two verbal idioms – *uzeti u obzir* 'take into consideration' and *voditi računa* 'take care'.

There are 533 synsets from the SrpWN which contain MWE literals that correspond to the synsets in the PWN used to annotate senses in ELEXIS-EN. Among them are 476 that were not annotated by our tools. Naturally, a number of them does not appear in ELEXIS-SR. For instance, *vodonik* 'hydrogen' appears in sentence 272, but its synonym *atomski broj 1* 'atomic number 1' does not. In some cases the Serbian correct translation avoids the use of a certain expression used in ELEXIS-EN, e.g. the sentence 597 ends with '...the best time being this summer.' while the same sentence in Serbian ends with *...najbolje ovog leta.* 'best this summer', avoiding the use of Serbian counter terms for 'time' in the sense 'a suitable moment', MWEs *pravi trenutak* or *pravi čas*. Finally, there are MWEs, like *merna jedinica* 'measurment unit', used in the sentence 735, which were not annotated since they were not yet recorded in lexical resources used for the annotation.

6. Linking MWEs and corpora

A holistic presentation of MWEs in lexicons and linking their entries with occurrences in a corpus is still an open question. We are considering the use of LLOD for interlinking MWE lexicon entries with their occurrences in corpora. Two options will be taken into account: Ontolex-lemon[7] (with Lexicog module) and DMLex[8]. Ontolex-lemon is widely used community standard for machine-readable lexical resources in the context of RDF, Linked Data, and Semantic Web technologies (McCrae et al., 2017). DMLex is a standard for structuring (human-oriented) dictionaries, which is published by LEXIDMA, a technical committee under OASIS, an organisation which oversees the production of open standards in the IT industry.

The following example gives an idea of how the lexical entry for MWE *fast food* can be represented in RDF along with its translations – in this case in Serbian *brza hrana* – and how links between senses and Wikidata entries are realized.

```
:le_fast_food
  a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "fast food"@en];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense
    [ontolex:reference
<https://www.wikidata.org/wiki/Q81799>];
  decomp:constituent :cm_fast;
  decomp:constituent :cm_food;
  rdf:_1 :le_fast; # lexical
  rdf:_2 :le_food. # entries

# component of cannonical form
:cm_food  a decomp:Component;
  decomp:correspondsTo :le_food.
  …
:le_brza_hrana a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "brza hrana"@sr];
  …

# simplified naming
:tranSetEN-SR vartrans:trans
 fast_food-ensns-brza_hrana-srsns .
:fast_food-ensns
 a ontolex:LexicalSense ;
 ontolex:isSenseOf :le_fast_food .
:brza_hrana-srsns
 a ontolex:LexicalSense ;
 ontolex:isSenseOf :le_brza_hrana .
:fast_food-ensns-brza_hrana-sns-trans
 a vartrans:Translation ;
 vartrans:source :fast_food-ensns ;
 vartrans:target :brza_hrana-srsns .
```

The OntoLex-FrAC[9] vocabulary implements the lexicon-corpus interface (Barbu Mititelu et al.,

[7] https://www.w3.org/2016/05/ontolex

[8] https://www.lexiconista.com/dmlex/ ;https://docs.oasis-open.org/lexidma/dmlex/v1.0/csd02/dmlex-v1.0-csd02.html

[9] The current draft version of the FrAC specification is found under https://github.com/ontolex/frequency-attestation-corpus-information/

2024) with corpus information to support corpus-driven lexicography and the inclusion of corpus evidence (attestations). A sentence number 823 in English: “It can be made at home or bought from fast food shops.” and in Serbian “Može se napraviti kod kuće ili kupiti u prodavnicama brze hrane.” illustrates this in the following example:

```
:le_fast_food
 frac:attestation [
 frac:quotation "It can be made at
 home or bought from fast food
 shops."@en;
 frac:observedIn :EWSD].

:le_brza_hrana
 frac:attestation [
 frac:quotation "Može se napraviti
 kod kuće ili kupiti u prodavnicama
 brze hrane."@sr;
 frac:observedIn :EWSD-ext].
```

The cross-lingual analysis of idiosyncratic constructions can be supported by publishing aligned and annotated corpus data as Linked Data employing community standards such as the NLP Interchange Format (NIF) (Hellmann et al., 2012) and CoNLL-RDF (Chiarcos and Fäth, 2017; Chiarcos and Glaser, 2020), a minimal NIF subset designed for compatibility with tab-separated formats used in NLP (“CoNLL”), Universal Dependencies (“CoNLL-U”) and Parseme (“Parseme-TSV”). The sense repository should be probably published using Ontolex-lemon. The first ideas about leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora are given in (Stanković et al., 2023).

7. Future Work

The development of the Serbian sentence set is a work in progress, as represented in Figure 5: translation and tokenization are done, POS tagging and lemmatization checking are in the final phase, and word sense inventory is being prepared. The syntactic annotation is still pending as well as the development of a LLOD dictionary.

Our future research will give special attention to the annotation of MWEs and NEs in the ELEXIS-SR. On the one hand, we will coordinate our work with the other research groups, primarily groups dealing with ELEXIS, Parseme, UD and UniDive activities. On the other hand, having in mind that the meticulously prepared set ELEXIS-SR will be used for various purposes, we plan to publish its various editions, e.g. annotating NEs using a large set of classes and sub-classes. One important research path will be the production of precise guidelines for distinguishing MWEs from NEs, as well as explicating differences in the notion of a MWE in the Serbian e-dictionaries, Parseme/UniDive and WordNet (e.g. MWEs *pravi trenutak* or *pravi čas* ‘time (a suitable moment)’ would probably not be considered a nominal MWE for Parseme/UniDive,[10] that is, they would not pass the prescribed sequence of tests).

The future research goal is the comparative analyses of MWEs and NEs in ELEXIS multilingual set both from the linguistic and NLP point of view.

Acknowledgements

This research was supported by the Ministry of Science, the Republic of Serbia, #GRANT 451-03-65/2024-03/ 200126,#GRANT 451-03-66/2024-03/200174 and COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive).

8. Bibliographical References

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Verginica Barbu Mititelu, Voula Giouli, Stella Markantonatou, Ivelina Stoyanova, Petya Osenova, Kilian Evang, Daniel Zeman, Simon Krek, Carole Tiberius, Christian Chiarcos, and Ranka Stankovic. 2024. Multiword expressions between the corpus and the lexicon: Universality, idiosyncrasy and the lexicon-corpus interface. In *Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)*.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2009. FrameNet for the semantic analysis of German: Annotation, representation and automation. *Multilingual FrameNets in Computational Lexicography: methods and applications*, 200:209–244.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017,*

[10]The guidelines for non-verbal MWEs are still in preparation.

Galway, Ireland, June 19-20, 2017, Proceedings 1, pages 74–88. Springer.

Christian Chiarcos and Luis Glaser. 2020. A tree extension for CoNLL-RDF. In *Proceedings of the 12th LREC*, pages 7161–7169.

Martelli Federico, Navigli Roberto, Krek Simon, Kallas Jelena, Gantar Polona, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. 2021. Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Proceedings of the eLex 2021 conference*, pages 377–395. Lexical Computing.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. NIF Combinator: Combining NLP Tool Output. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, 2012. Proceedings 18*, pages 446–449. Springer.

Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Construction of large-scale English verbal multiword expression annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Cvetana Krstev. 2021. White as snow, black as night – similes in old serbian literary texts. *Infotheca - Journal for Digital Humanities*, 21(2):119–135.

Cvetana Krstev, Ivan Obradović, Ranka Stanković, and Duško Vitas. 2013. An approach to efficient processing of multi-word units. *Computational Linguistics: Applications*, pages 109–129.

Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2):473–489.

Cvetana Krstev, Ranka Stanković, and Aleksandra Marković. 2023. Multiword expressions – comparative analysis based on aligned corpora. In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France*.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. Building semantic concordances. In Christian Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Bolette Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023a. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Bolette Sandford Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023b. The DA-ELEXIS Corpus – a Sense-Annotated Corpus for Danish with Parallel Annotations for Nine European Languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, and et al. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, et al. 2018. PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5905–5911.

Ranka Stanković, 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, Christian Chiarcos, and Milica Ikonić Nešić. 2023. Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora. In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France*.

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 Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114, Sofia, Bulgaria. Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences.

Ranka Stanković, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. Parallel Bidirectionally Pretrained Taggers as Feature Generators. *Applied Sciences*, 12(10).

Ralph Weischedel, Eduard Hovy, Marcus Mitchell, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural LanguageProcessing and Machine Translation: DARPA GlobalAutonomous Language Exploitation*.

Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. 2019. Development and evaluation of three named entity recognition systems for serbian - the case of personal names. In *Natural Language Processing in a Deep Learning World – Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1061–1068. INCOMA Ltd.

9. Language Resource References

Martelli, Federico and Navigli, Roberto and Krek, Simon and Kallas, Jelena and Gantar, Polona and Koeva, Svetla and Nimb, Sanni and Sandford Pedersen, Bolette and Olsen, Sussi and Langemets, Margit and Koppel, Kristina and Üksik, Tiiu and Dobrovoljc, Kaja and Ureña-Ruiz, Rafael and Sancho-Sánchez, José-Luis and Lipp, Veronika and Váradi, Tamás and Győrffy, András and Simon, László and Quochi, Valeria and Monachini, Monica and Frontini, Francesca and Tiberius, Carole and Tempelaars, Rob and Costa, Rute and Salgado, Ana and Čibej, Jaka 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. 2023. *Parallel sense-annotated corpus ELEXIS-WSD 1.1*. The Jožef Stefan Institute. Slovenian language resource repository CLARIN.SI.