Annotation of MWEs and NEs in the Serbian extension of ELEXIS-WSD: comparisons, solutions and open questions

Cvetana Krstev
Association for Language Resources and Technologies
Belgrade, Serbia
cvetana@jerteh.rs

Ranka Stanković
University of Belgrade
F. of Mining and Geology
Belgrade, Serbia
ranka@rgf.bg.ac.rs

Aleksandra Marković
Institute for the Serbian Language SASA
Belgrade, Serbia
malexa39@gmail.com

Relevant UniDive working groups: WG2

1 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).[1] The list of sense inventories is based on WordNet for DA (Pedersen et al., 2023), 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[2] in EN was translated automatically (Google translation) into SR. A few (eight precisely) Serbian native speakers checked the translation, and after that sentences were read carefully in order to resolve different issues: 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 in order to choose the right morphological form); besides, it was necessary to check phonetic transcriptions of names (particularly personal ones), since in SR it is not usual to write names in the original form. 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, their content is of encyclopedic sort, and often it was necessary to read the original document in English or some other language to understand the meaning and represent it in SR.

After this process, the set was proofread and automatically tokenized, lemmatized, and POS-tagged (Stankovic et al., 2020; Stanković et al., 2022). The outcomes of all these automatic procedures were manually corrected. Tasks that remain to be done include the annotation of MWEs and NEs, the syntactic annotation, and linking with the sense repository.

Each language has a separate sense inventory containing all senses (and their definitions) used for the annotation in the corpus, but not all the senses from the sense inventory are necessarily included in the corpus annotations.

2 MWEs and NEs in WSD

In this paper, we are focusing on the annotation of MWEs and NEs (see Table 1). The second and fourth columns present the number of unique lemmas in the WSD, while the third and fifth present the number of unique senses.

	MWE		NE	
Lang.	**lemma**	**sense**	**lemma**	**sense**
bg	299	465	2	2
da	440	477	440	459
en	179	309	1	1
es	36	40	4	8
et	177	217	112	145
hu	7	7	6	6
it	41	42	0	0
nl	33	37	27	27
pt	113	115	14	15
sl	385	451	0	0
Total	1,710	2,160	606	663

Table 1: Number of MWEs/NEs in the repository.

All MWEs and NEs occurring in the whole repository were automatically translated into SR (as phrases, not word-to-word) and the number of translation equivalents that were exact

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

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

matches was calculated. Table 2 shows that from a total of 1,412 translations, one international MWE appeared in 6 language sets, *lingua franca*. One of two MWEs translated from 4 languages (5 senses) into one SR term was *средња школа* (SR): *висше училище* (BG), *high school* (EN), *srednja šola* (SL), *visoka šola* (SL), *scuola media* (IT).[3] Named entities were not systematically introduced for all languages, resulting in a total of 526 entries. The most frequent NE was *Грчка*, translated from four languages: *Grækenland* (DA), *Grecia* (ES), *Kreeka* (ET), *Grécia* (PT).

№	1	2	3	4	5	6	Tot.
MWE	92	163	40	14	2	1	1412
NE	453	67	5	1	0	0	526

Table 2: MWEs and NEs translations into Serbian obtained by translating from 1 to 6 languages.

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

(1) The dictionary of non-verbal MWEs was used to annotate this type of MWEs (653 occurrences) (Krstev et al., 2013). Among them were 357 nominal MWEs, 134 proper nouns, 73 prepositions, 52 adverbs, 36 conjunctions, and one adjective.

(2) A system for the Named Entity Recognition (NER) based on e-dictionaries and rules (2,006 occurrences) (Krstev et al., 2014). Numbers of recognized NEs per class are presented in Table 3. Some MWEs are recognized both by dictionaries and the NER system.

(3) 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) (228 occurrences, distribution by type: IRV – 174, LVC.full – 35, VID – 11, and LVC.cause – 8).

(4) A system for the recognition of adjectival and verbal similes described in (Krstev et al., 2023). Not a single simile was retrieved in this set of sentences.

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

The accuracy of MWE/NE recognition, recall and precision will be determined during the next step, when senses will be associated to simple- and multi-word units.

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 3: Recognized NEs by type.

3 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 (SSS) was based on their automatic translation to SR, as explained in Section 2. This was not ideal, since in a number of cases the translation was not precise: e.g., the SR verb *ustati* was obtained as a translation equivalent of four VMWEs from three languages, all with different meanings and scattered across unrelated sentences: *издигам се* (BG: 475, 792), *holde stand* (DA: 1363), *stå op* (DA: 856, 1244), *üles kasvatama* (ET: 1168). Verbs corresponding to those VMWEs in SSS were also various: *nastati* (475), *izdizati* (792), *izlaziti* (856, 1244), *odgajati* (1168), the translated verb *ustati* wasn't among them. On the other hand, in many cases matches were good: e.g., the SR translation *teška voda* 'heavy water' was obtained from MWEs in four languages: *tungt vand* (DA), *agua pesada* (ES), *zwaar water* (NL), *težka voda* (SL), all occurring in the same sentence – 1642. In the corresponding sentence in SSS the translated term *teška voda* was used and annotated as MWE.

In other cases, the translation was good, it was used in SSS, but it was not annotated in it because it was missing in the used resources. This was the case for *društvena mreža* 'social network', translated from *социална мрежа* (BG), *rede social* (PT), *družabno omrežje*, *družbeno omrežje* (SL), used twice in SSS, but not annotated. This case of missing annotations occurs in other languages as well. E.g., equivalents for *teška voda – heavy water* (EN),

тежска вода (BG) – were not annotated as MWE; *acqua pesante* (IT), *raske vesi* (ET) were annotated as MWE but not translated as *teška voda*, while *água-pesada* (PT) and *nehézvizet* (HU) were single tokens.

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 SSS, 116 MWE lemmas occurred in at least one language set in WSD; out of 228 VMWE occurrences (99 lemmas) annotated in SSS, 11 lemmas occurred in at least one language set; only 93 NEs annotated in SSS were annotated as MWE or PROPN in WSD (maybe due to the poor lemmatization and linking of proper names).

4 Open questions

The development of the Serbian sentence set is a work in progress: translation and tokenization are done, presently POS tagging and lemmatization are being checked and corrected, and word sense inventory is being prepared. MWE and NE tagging will follow decisions taken by the UniDive action.

Open questions for future work, concerning SSS, but other languages as well, are: (a) finding the alternative ways of aligning MWEs and NEs across languages (Ivačič et al., 2023); (b) should the set of sentences be enhanced to capture a more versatile style, e.g. fiction (as the lack of simile figures suggests)? (c) should the repertoire of NE classes be unique for all languages? (d) should NEs include numeric and/or temporal expressions? (e) should the nesting of MWEs/NEs be allowed?

Acknowledgements

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

References

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.

Nikola Ivačič, Thi Hong Hanh Tran, Boshko Koloski, Senja Pollak, and Matthew Purver. 2023. Analysis of transfer learning for named entity recognition in south-slavic languages. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 106–112.

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.*

Bolette Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023. 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.

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.

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, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. Parallel bidirectionally pretrained taggers as feature generators. *Applied Sciences*, 12(10).