



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

Cvetana Krstev⁽¹⁾, Ranka Stanković⁽²⁾, Aleksandra Marković⁽³⁾

1) Jerteh; 2) UB FMG; 3) SASA ISL; Belgrade, Serbia

1. The extension of ELEXIS-WSD

ELEXIS-WSD – parallel sense-annotated corpus; N, Adj, V, Adv

with assigned senses for 10 languages: BG, DA, EN, ES, ET, HU, IT, NL, PT, SL.

The set of 2024 sentences from WikiMatrix in EN was translated automatically into SR; the translation was

checked by native speakers and linguists and proofread afterwards.

1	RB	En	Sr	Google translation	Corr (ovde stavljati korigovane vrednosti, da original ipak ostane) Pismo, ćirilica
1584	1583.en	The gas is heavier than air.	Течности су вискозније од гасова.	Гас је тежи од ваздуха.	Гас је тежи од ваздуха.
1585	1584.en	The strategic position of the Ardennes has made it a battleground for European powers for centuries.	/	Стратешки положај Ардена чинио их је вековима бојним пољем европских сила.	Стратешки положај који су Ардени имали чинио их је током векова поприштем европских сила.
1586	1585.en	He also established the relationship between barometric pressure and height above sea level.	/	Такође је установио везу између барометарског притиска и висине изнад нивоа мора.	Установио је и везу између барометарског притиска и надморске висине.

1	Rece	word	tag_p	tag_u	lemma	komenar	POS
9539	1583	je	V	AUX	jesam		
9540	1583	тежи	V	VERB	težiti		
9541	1583	од	PREP	ADP	od		
9542	1583	ваздуха	N	NOUN	vazduh		
9543	1583	.	SENT	PUNCT	.		
9544	1583	</s>					
9545	1584	<s>					
9546	1584	Стратешки	A	ADJ	strateški		
9547	1584	положај	N	NOUN	položaj		
9548	1584	који	PRO	DET	koji		
9549	1584	Ардени	N	NOUN	Ardeni		N/PROPN
9550	1584	имају	V	VERB	imati		
9551	1584	чинио	V	VERB	činiti		
9552	1584	их	PRO	PRON	ona		
9553	1584	је	V	AUX	jesam		
9554	1584	током	PREP	ADP	tokom		
9555	1584	векова	N	NOUN	vek		
9556	1584	поприштем	N	NOUN	poprište		
9557	1584	европских	A	ADJ	evropski		
9558	1584	сила	N	NOUN	sila		
9559	1584	.	SENT	PUNCT	.		
9560	1584	</s>					
9561	1585	<s>					
9562	1585	Установио	V	VERB	ustanoviti		
9563	1585	је	V	AUX	jesam		
9564	1585	и	CONJ	PART	i		
9565	1585	везу	V	VERB	vesti		
9566	1585	између	PREP	ADP	između		
9567	1585	барометарског	A	ADJ	barometarsk		
9568	1585	притиска	N	NOUN	pritsak		
9569	1585	и	CONJ	CCONJ	i		
9570	1585	надморске	A	ADJ	nadmorski		
9571	1585	висине	N	NOUN	visina		
9572	1585	.	SENT	PUNCT	.		
9573	1585	</s>					

2. MWEs & NEs in WSD

Set automatically tokenized, lemmatized, POS-tagged, and manually corrected.

Lang.	MWE		NE	
	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

All MWEs & NEs from the WSD automatically translated into SR.



Next steps:

- MWEs & NEs annotation
- syntactic annotation
- linking with the sense repository.

Number of exact matches:

MWEs:

- *lingua franca* appeared in 6 language sets;
- *high school* SR: *srednja škola* (lit. 'middle school'); BG: *вучие училище*; EN: *high school*; SL: *srednja šola* & **visoka šola*; IT: *scuola media*

NEs: *Grčka* 'Greece' was the most frequent NE, translated from: *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.

Correct equivalents:

<i>ustati</i> (SR) 'get up' as incorrect translation of:	----->	<i>nastati</i> 'become'
<i>uzdugam se</i> (BG)	----->	<i>izdizati</i> 'elevate'
<i>holde stand</i> (DA)	----->	<i>izlaziti</i> 'rise'
<i>stå op</i> (DA)	----->	<i>odgajati</i> 'raise'
<i>üles kasvatama</i> (ET)	----->	

In some cases the translation of MWEs was good and used in Serbian set, but not annotated in it because it was missing in the above resources: *društvena mreža* 'social network'.

The same holds for other languages: *heavy water* (EN) & *тежка вода* (BG) were not annotated neither.

3. The comparison of MWEs & NEs across languages

Automatic translation of MWEs sometimes imprecise:

ustati (SR) 'get up' as incorrect translation of:

<i>ustati</i> (SR) 'get up' as incorrect translation of:	----->	<i>nastati</i> 'become'
<i>uzdugam se</i> (BG)	----->	<i>izdizati</i> 'elevate'
<i>holde stand</i> (DA)	----->	<i>izlaziti</i> 'rise'
<i>stå op</i> (DA)	----->	<i>odgajati</i> 'raise'
<i>üles kasvatama</i> (ET)	----->	

4. Open questions & future work

- (a) finding the alternative ways of aligning MWEs and NEs across languages;
- (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?

Serbian set automatically annotated using 4 resources & tools:

(1) Dictionary of non-verbal MWEs (653 occurrences): 357 NMWEs (134 PROPN), 73 PREP, 52 ADV, 36 CONJ, 1 ADJ.

(2) NER based on e-dictionaries and rules (2006 occurrences).

(3) Recognition of VMWEs based on e-dictionaries, rules, and the repertoire of VMWEs annotated in the Serbian part of the PARSEME 1.3 (228 occurrences: IRV – 174, LVC.full – 35, VID – 11, and LVC.cause – 8).

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.

(4) Recognition of adjectival and verbal similes: not a single simile was retrieved.

The equivalents in some languages are single tokens: *águá-pesada* (PT) and *nehézvizet* (HU).

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).

