Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
LREC–COLING 2024, 25 May, Turin, Italy

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

Cvetana Krstev, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov

УНИВЕРЗИТЕТ У БЕОГРАДУ-РУДАРСКО-ГЕОЛОШКИ ФАКУЛТЕТ

Институт за српски језик САНУ

What will we talk about

1. The work on ELEXIS-sr corpus – recognition of MWEs & NEs; comparison with annotations in other languages
2. Building the Serbian sense inventory (Serbian WordNet)
3. Dictionaries of MWEs as LLOD, linking entries with their occurencies in the corpus

The extension of ELEXIS-WSD – ELEXIS-SR

Done so far (Krstev et al., 2024):

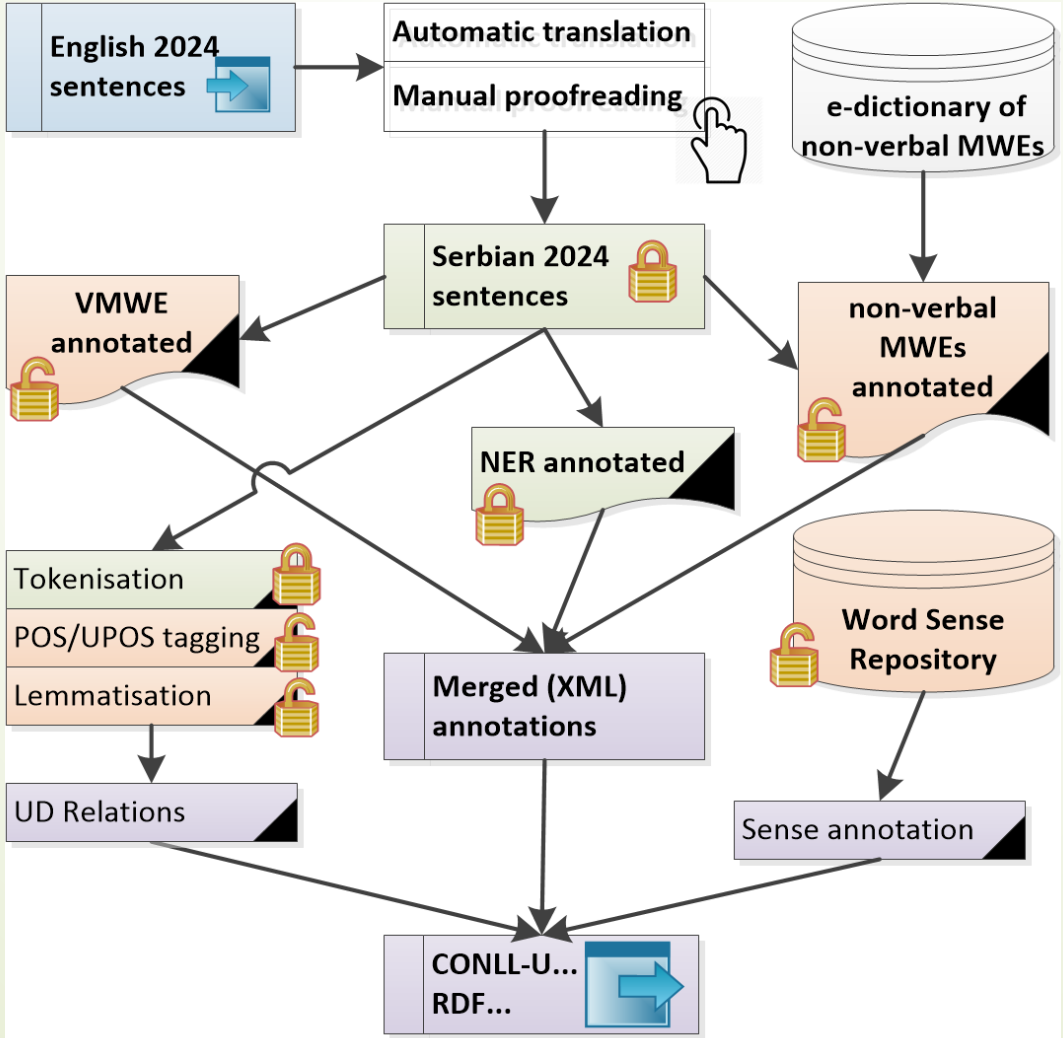
- automatically translated EN SS (sentence set), manually checked, proofred; reference resolution; phonetic transcritpion of NEs
- SR SS automatically tokenized, lemmatized, POS-tagged (Stanković et al., 2020; Stanković et al., 2022); manual correction in the final phase

The extension of ELEXIS-WSD – ELEXIS-sr

To be done:

- annotation of MWEs and NEs
- finishing the sense repository (SrpWN)
- semantic annotation
- syntactic annotation

Annotation of MWEs and NEs in the Serbian dataset

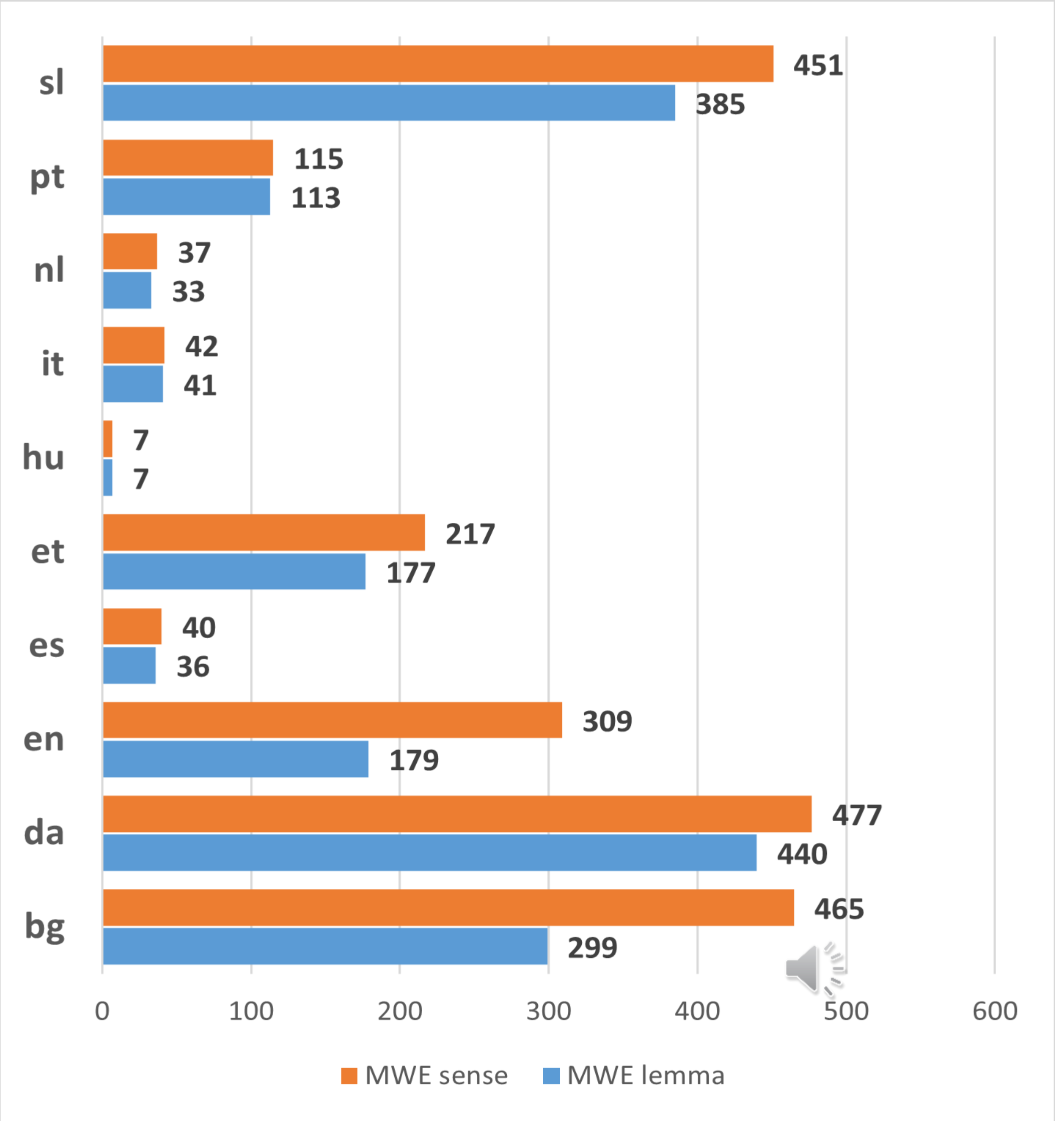


MWEs and NEs in WSD

- uneven number of MWEs and NEs annotated in 10 language SSs (due to different resources)
- MWEs and NEs from ELEXIS-WSD automatically translated into SR (as phrases) in order to facilitate the comparison with MWEs/NEs retrieved in the SR SS
- MWE in one language may be translated as a single word: *prime minister* (EN) → *premijer* (SR)

MWEs in ELEXIS-WSD

1,710 MWEs in 10 lang.; the numbers of MWEs differ significantly between languages



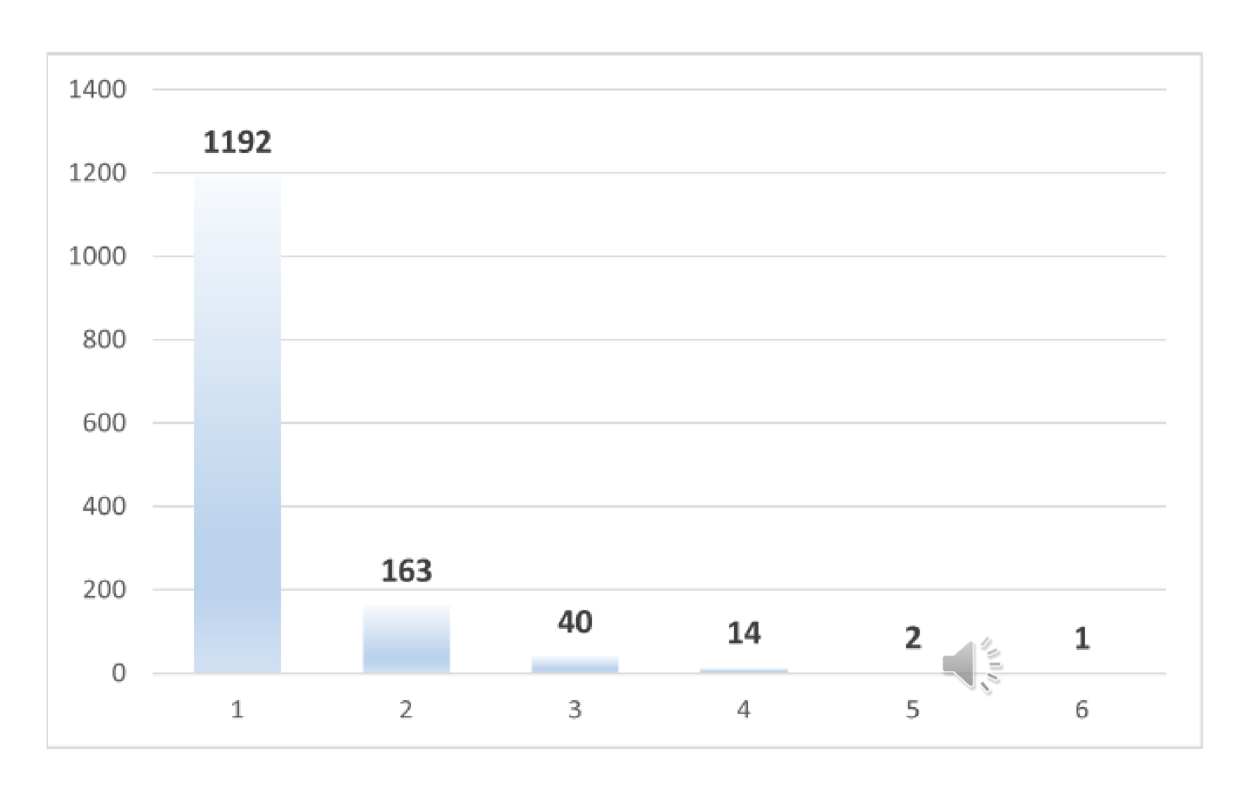
MWEs in ELEXIS-WSD

1) *Lingua franca* (6)

2) 14 MWEs (4):

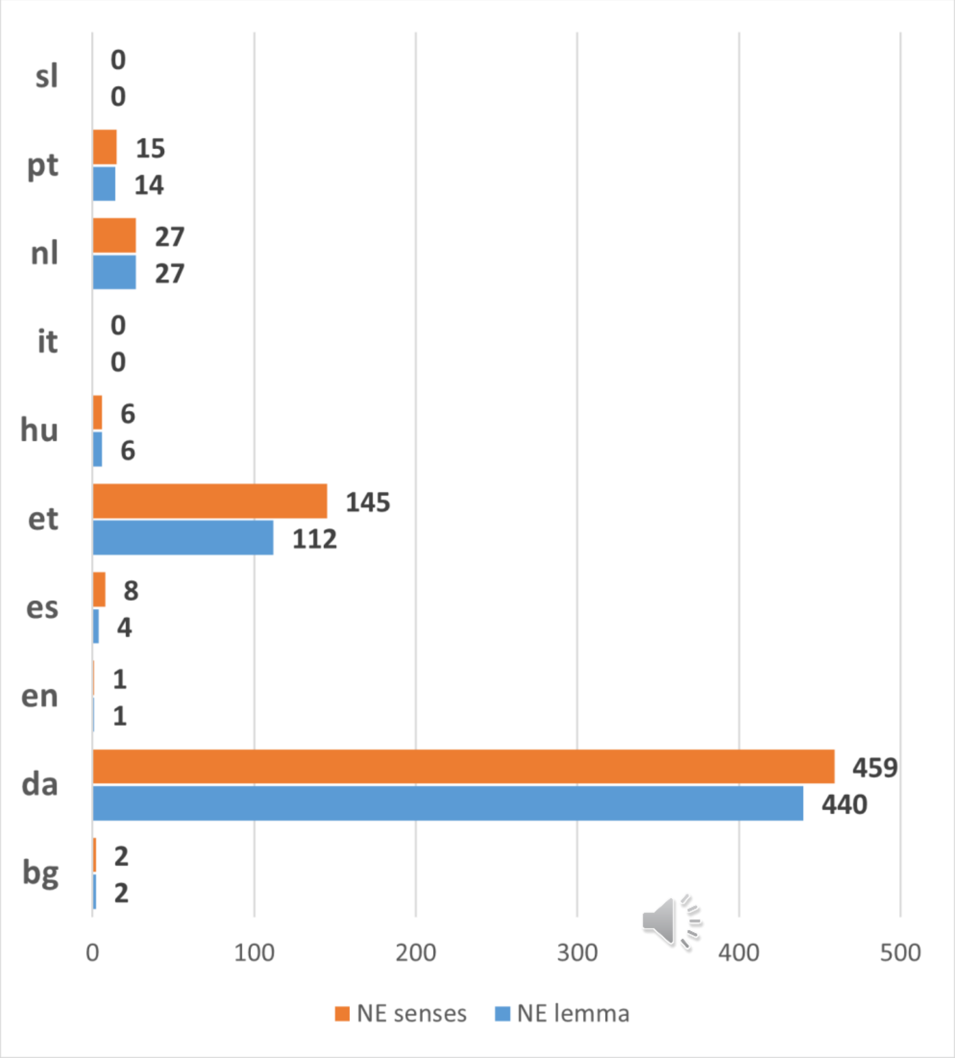
- *Life expectancy* (EN)
- *expectativa de vida* (PT)
- *pričakovana življenjska doba* (SL)
- *oodatav eluiga* (ET)

1,710 MWEs → 1,412 translations

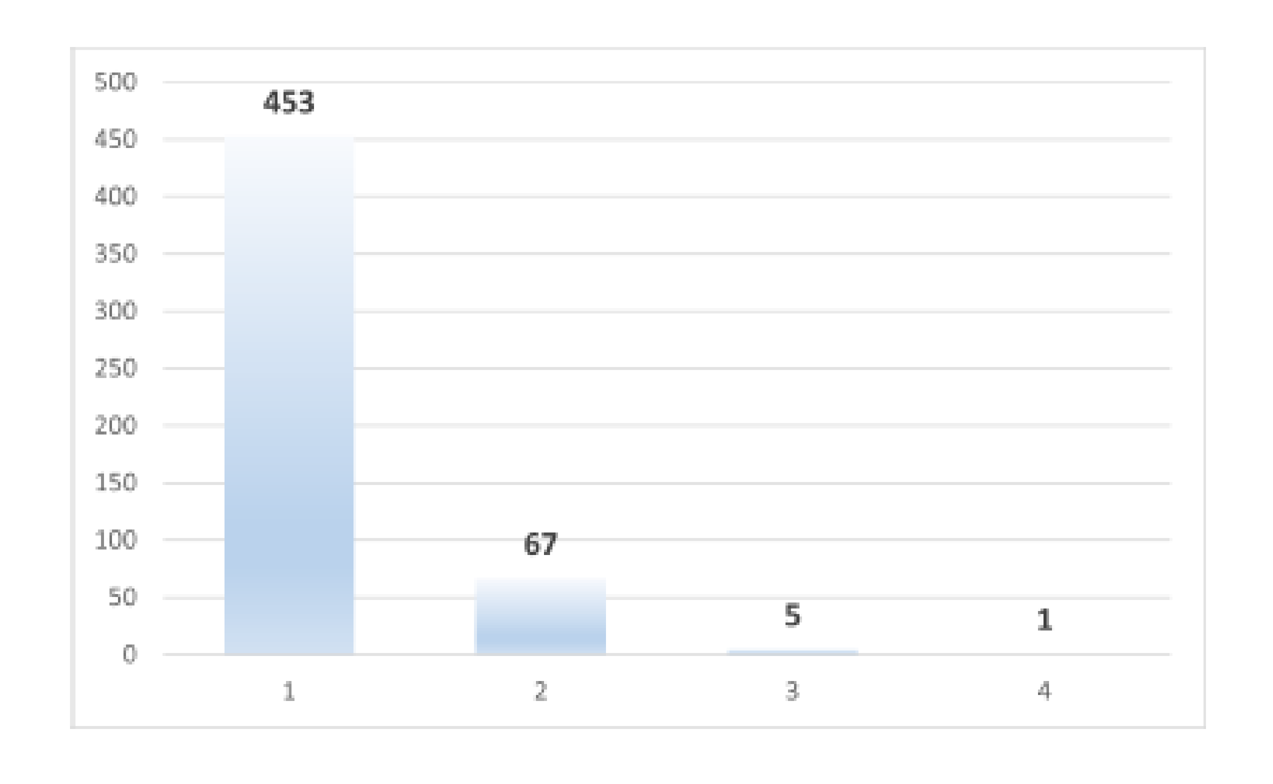


NEs in ELEXIS-WSD

- 606 NEs were annotated in 10 lang. SSs (number of NEs annotated is uneven)
- 526 translations, the most frequent: *Grækenland* (DA), *Grecia* (ES), *Kreeka* (ET), *Grécia* (PT) (from 4 lang.)



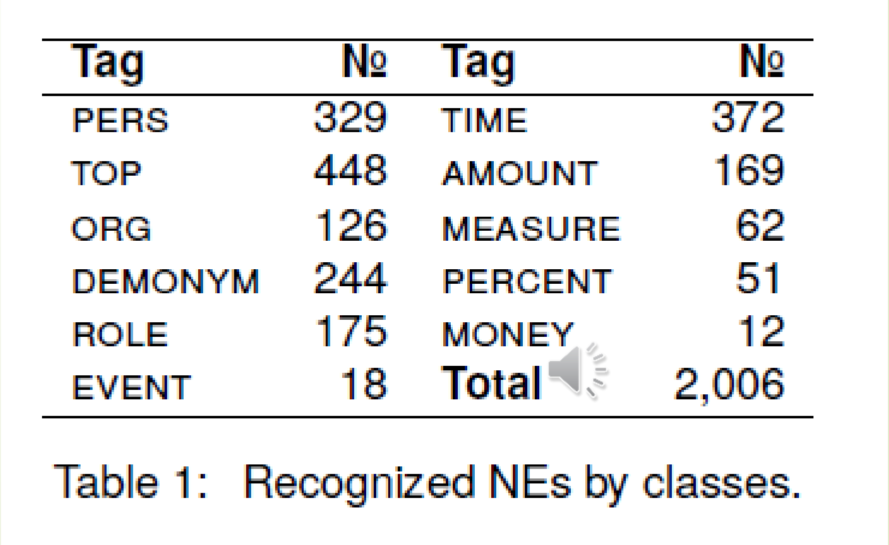
NEs in ELEXIS-WSD



Tools & resources used in automatic annotation of the Serbian SS

1. The e-dictionary of non-verbal MWEs (Krstev et al., 2013): 529 annotations (339 different) (Krstev *et al.*, 2013).

2. A system for NER (Krstev *et al.*, 2014): 2,006 occurrences of NEs.



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.

Tools & resources used in automatic annotation of the Serbian SS

3. A system for the recognition of VMWEs Serbian part of the PARSEME Corpus Release 1.3 (Savary et al., 2023) (230 VMWEs, 98 diff.)

4. A system for the recognition of adjectival and verbal similes (Krstev et al., 2023) (zero similes).

The comparison of MWEs and NEs across languages

- [1258] *bruto domaći proizvod* (SR) ‘gross domestic product’ (annotated as MWE)
- *gross domestic product* (EN), *produto interno bruto*(PT), *bruto [domači proizvod]* (SL), *sisemajandus koguprodukt* (ET) automatically translated into SR & annotated as MWE
- *bruttonationalproduktet* (DA); PIL (IT)

The comparison of MWEs and NEs across languages

- in some cases the automatic translation was good and it was used in ELEXIS-SR, but not annotated (due to incomplete resources)
- [1560] *prirodna selekcija* (SR), *natural selection* (EN), *seleção natural* (PT), *naravni izbor* (SL)
- missed annotation also in BG, ES, HU, NL

The comparison of MWEs and NEs across languages – comparison

ELEXIS-sr	WSD
653 non-V MWEs (384 lemmas)	116 MWE lemmas in at least 1 LS
228 VMWE (99 lemmas)	11 lemmas in at least 1 LS
93 NEs annotated (2,006)	MWE/PROPN

Sense repository

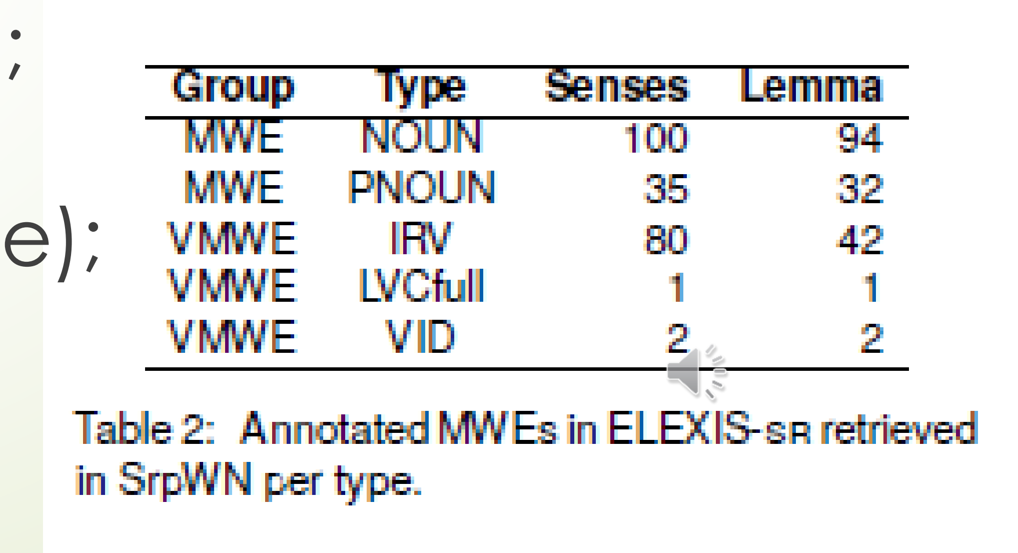
- Serbian WordNet (SrpWN) (Stanković et al. 2018) (25,322 synsets)
- ELEXIS-EN sense repository (based on PWN) 16,106 entries; aligned with PWN synsets → 13,703 matches
- Synsets from this subset aligned with SrpWN → 5,997 matches

Sense repository

- The subset missing from the list of 13,703 synsets compared with sentence annotations in ELEXIS-EN – a gap of 2,130 synsets
- automatic translation of synsets and their definitions from the PWN (Google API & OpenAI services)
- Ongoing: postediting the list of synonym set candidates & definitions

Sense repository

- 437 MWEs annotated in ELEXIS-SR (339 non-V & 98 VMWEs) 171 (39%) found in SrpWN
- polysemous MWEs; IRV *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).



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.

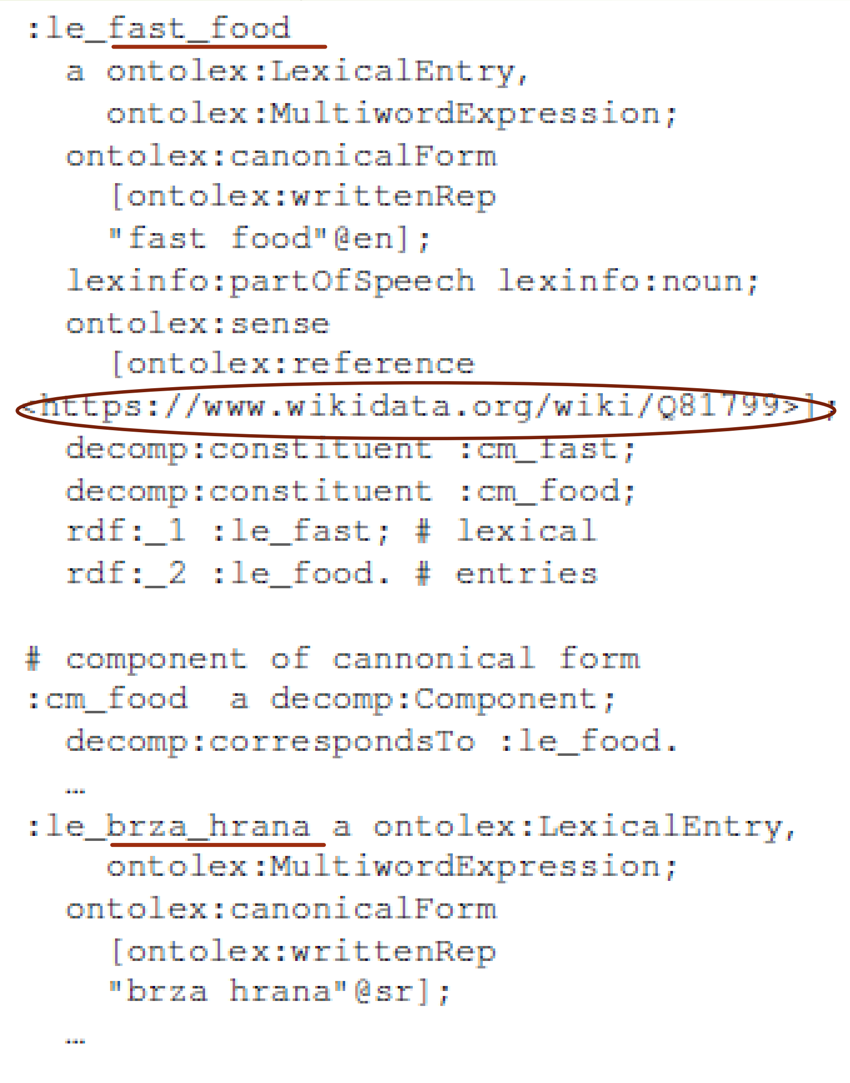
Linking MWEs and corpora

- Open issue: encoding MWEs in lexicons and linking entries with occurences in corpus
- LLOD for interlinking: 1) Ontolex-lemon (Lexicog module) 2) DMLex

Ontolex-lemon – standard for machine-readable lex. resources (McCrae et al., 2017)

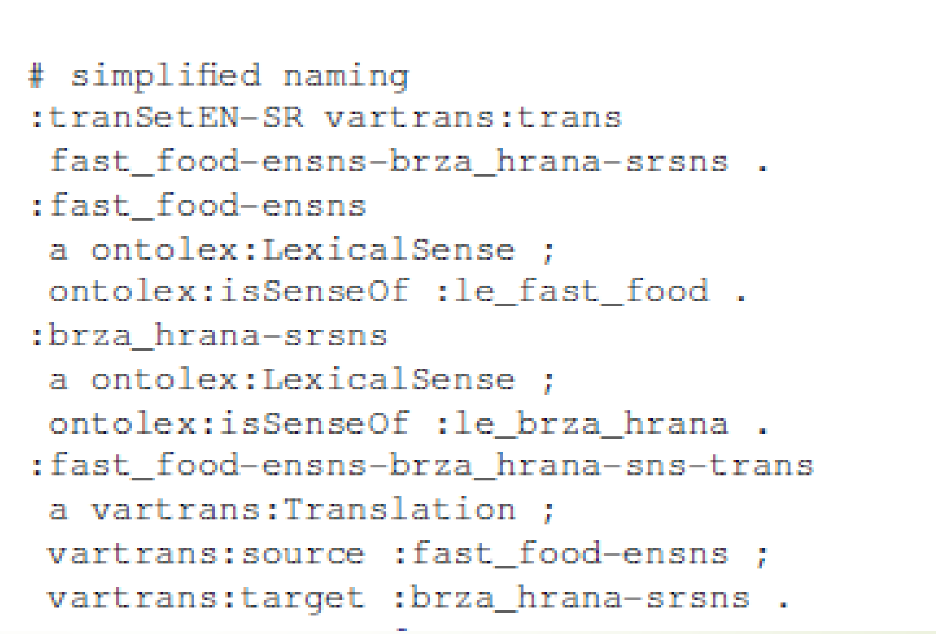
DMLex – standard for structuring dictionaries (LEXIDMA)

Linking MWEs and corpora



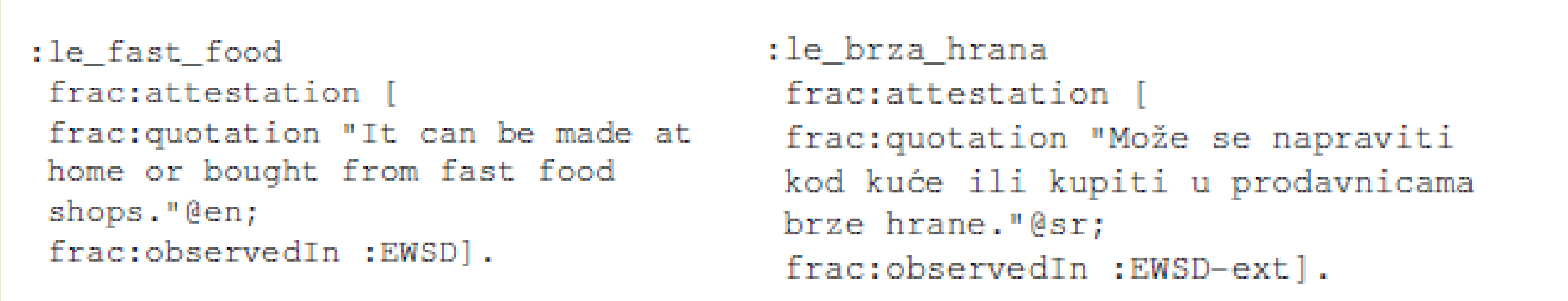
```
:le_fast_food
  a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "fast food"@en];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense
    [ontolex:reference
<https://www.wikidata.org/wiki/Q81799>];
  decomp:constituent :cm_fast;
  decomp:constituent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

Linking MWEs and corpora



```
: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].
```

(Barbu-Mititelu et al., 2024)

[823] “It can be made at home or bought from fast food shops.”

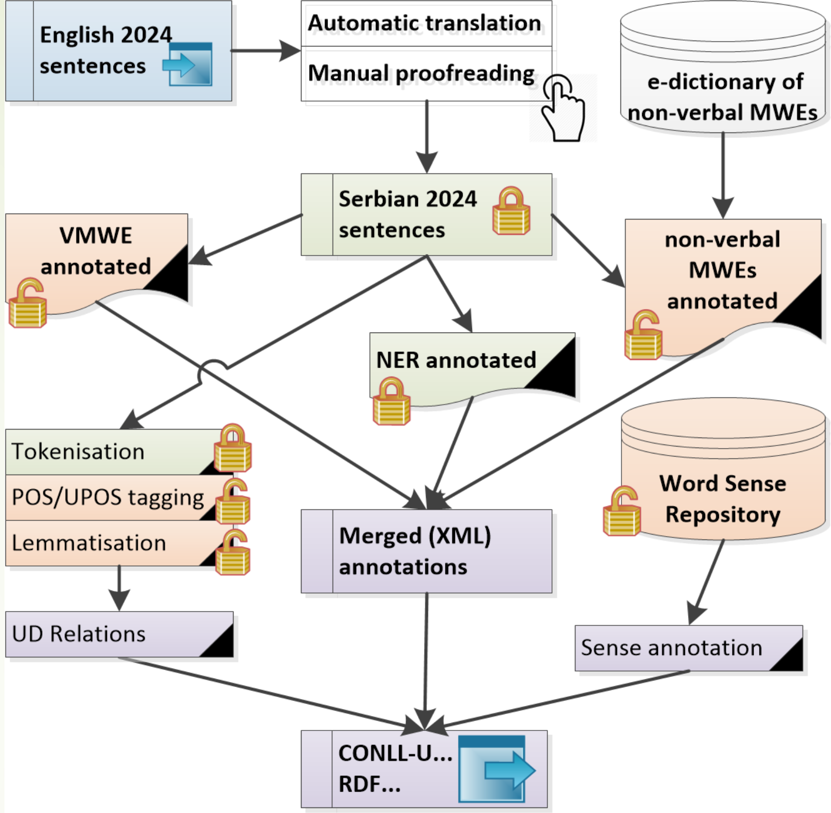
„Može se napraviti kod kuće ili kupiti u prodavnicama brze hrane.“

Linking MWEs and corpora

- Publishing aligned and annotated corpus data as Linked Data: NIF (Helman et al., 2012) & CoNLL-RDF (Chiarcos and Fäth, 2017; Chiarcos and Glaser, 2020).
- Ontolex-lemon for publishing sense repository.
- Stanković et al., 2023.

Future work

1. finished: translation&tokenization
2. final phase: POS-tagging, lemmatisation checking;
3. preparing of sense inventory;
4. pending: syntactic annotation & LLOD dictionary



Future work

- annotation of MWEs & NEs in ELEXIS-SR
- coordination with other research groups (ELEXIS, Parseme, UD, UniDive)
- precise guidelines for distinguishing MWEs from NEs
- differences in the notion of MWE in the SR e-dictionaries, Parseme/UniDive & WN
- comparative analysis of MWEs&NEs in ELEXIS multilingual set from the linguistic &NLP point of view

Thanks for listening!