



Annotation of Multiword Expressions in the SUK 1.0 Training Corpus of Slovene: Lessons Learned and Future Steps

Jaka Čibej, Polona Gantar, Mija Bon

Centre for Language Resources and Technologies,
University of Ljubljana

Joint Workshop on Multiword Expressions and Universal
Dependencies (MWE-UD 2024) @ LREC-COLING 2024

Torino, Italy, 25 May 2024

Motivation

- **SUK 1.0 Slovene Training Corpus**
 - Verbal MWEs in PARSEME (~11.500 sentences)
 - after PARSEME: non-verbal MWEs (~6.500 sentences) – proof-of-concept
 - non-verbal MWE annotations not finalized and not included in the corpus yet
- **Continuation within UniDive**
 - insight for MWE guidelines
 - finalization of annotations
 - adding labels according to new PARSEME categories
 - quantitative analysis: POS structures, scope of MWE annotation; overlap with named entities

Task

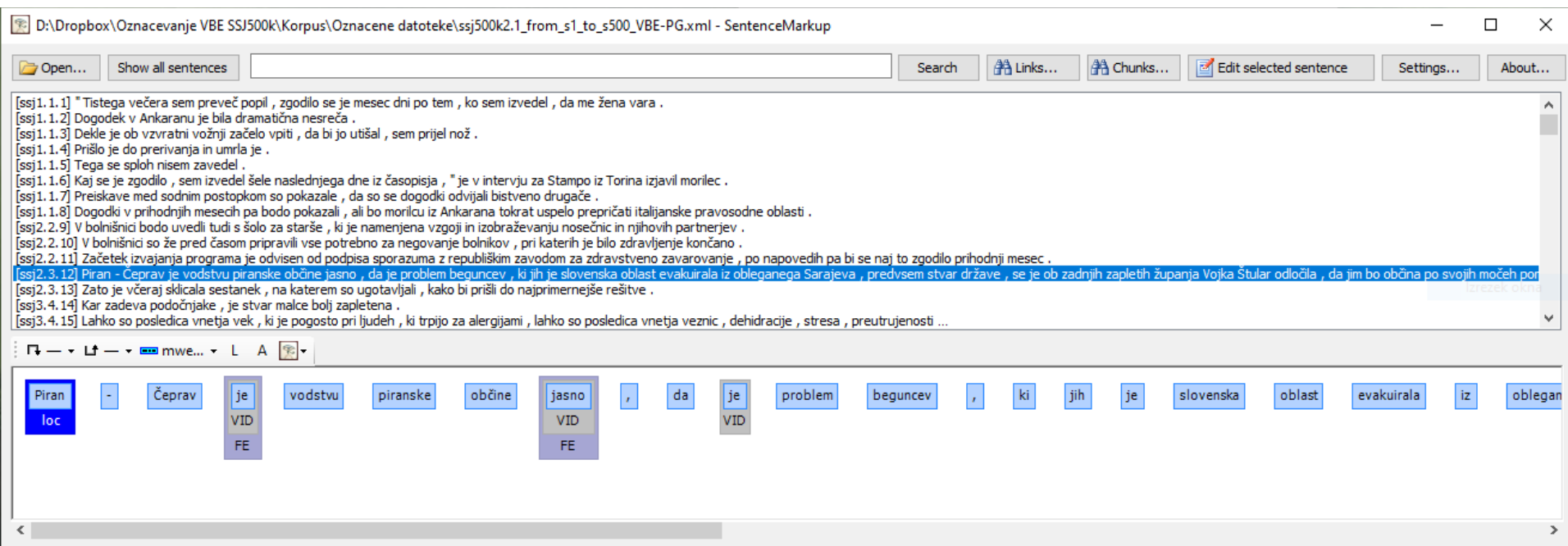
- **Identification and categorization of non-verbal MWEs**
 - from the point of view of their inclusion in dictionaries!
 - meaning + syntactic function
 - **fixed expressions** (*črna luknja* ‘black hole’)
 - **phraseological units** (*princ na belem konju* lit. ‘prince on a white horse’; ‘knight in shining armor’)
 - **syntactic combinations** (*v skladu z* ‘in accordance with’)

Annotation Software

- **Q-CAT (Querying-Supported Corpus Annotation Tool)**

<http://hdl.handle.net/11356/1844>

- open-access; supports COnLL-U
- custom annotation layers
- offline; single annotator per file



D:\Dropbox\Oznacevanje VBE S5J500k\Korpus\Oznacene datoteke\ssj500k2.1_from_s1_to_s500_VBE-PG.xml - SentenceMarkup

Open... Show all sentences Search Links... Chunks... Edit selected sentence Settings... About...

[ssj1.1.1] * Tistega večera sem preveč popil, zgodilo se je mesec dni po tem, ko sem izvedel, da me žena vara .
 [ssj1.1.2] Dogodek v Ankaranu je bila dramatična nesreča .
 [ssj1.1.3] Dekle je ob vzratni vožnji začelo vpiti, da bi jo utišal, sem prijel nož .
 [ssj1.1.4] Prišlo je do prerivanja in umrla je .
 [ssj1.1.5] Teга se sploh nisem zavedel .
 [ssj1.1.6] Kaj se je zgodilo, sem izvedel šele naslednjega dne iz časopisja, " je v intervju za Stampo iz Torina izjavil morilec .
 [ssj1.1.7] Preiskave med sodnim postopkom so pokazale, da so se dogodki odvijali bistveno drugače .
 [ssj1.1.8] Dogodki v prihodnjih mesecih pa bodo pokazali, ali bo morilu iz Ankarana tokrat uspelo prepričati italijanske pravosodne oblasti .
 [ssj2.2.9] V bolnišnici bodo uvedli tudi s šolo za starše, ki je namenjena vzgoji in izobraževanju nosečnic in njihovih partnerjev .
 [ssj2.2.10] V bolnišnici so že pred časom pripravili vse potrebno za negovanje bolnikov, pri katerih je bilo zdravljenje končano .
 [ssj2.2.11] Začetek izvajanja programa je odvisen od podpisa sporazuma z republiškim zavodom za zdravstveno zavarovanje, po napovedih pa bi se naj to zgodilo prihodnji mesec .
 [ssj2.3.12] Piran - Čeprav je vodstvu piranske občine jasno, da je problem beguncev, ki jih je slovenska oblast evakuirala iz obleganega Sarajeva, predvsem stvar države, se je ob zadnjih zapletih županja Vojka Štular odločila, da jim bo občina po svojih močeh por...
 [ssj2.3.13] Zato je včeraj sklicala sestanek, na katerem so ugotavljali, kako bi prišli do najprimernejše rešitve .
 [ssj3.4.14] Kar zadeva podočnjake, je stvar malce bolj zapletena .
 [ssj3.4.15] Lahko so posledica vnetja vek, ki je pogosto pri ljudeh, ki trpijo za alergijami, lahko so posledica vnetja veznic, dehidracije, stresa, preutrujenosti ...

Piran - Čeprav je vodstvu piranske občine jasno, da je problem beguncev, ki jih je slovenska oblast evakuirala iz obleganega Sarajeva, predvsem stvar države, se je ob zadnjih zapletih županja Vojka Štular odločila, da jim bo občina po svojih močeh por...

loc VID FE VID VID

Results

- 10 annotators (2 reference annotators + 8 students)
- **15,727** MWE annotations in the first **6,500** sentences of SUK 1.0.
- Each sentence was annotated by at least 3 annotators.
- **8,864** MWE candidates in total
- **6,385** different potential MWEs

- The corpus ALSO contains manual UD POS tags, UD dependency relations, named entity annotations
 - cross-reference MWE annotations with other annotation layers
 - identify patterns and points of disagreement

MWE Annotations by POS-structure

- 920 different structures, with the top 17 accounting for approx. 65% of all annotations
- mostly non-verbal MWEs
- **ADJ NOUN** (sodni postopek, ‘judicial process’)
- **ADP NOUN** (v celoti, lit. ‘in whole’, ‘entirely’)

Structure	MWE Ann.	%
ADJ NOUN	4,550	29.04%
ADP NOUN	2,053	13.10%
ADP DET	401	2.56%
NOUN NOUN	391	2.50%
VERB ADP NOUN	360	2.30%
PART AUX	353	2.25%
ADP DET NOUN	298	1.90%
PART ADV	228	1.46%
ADJ ADJ NOUN	224	1.43%
ADP ADJ NOUN	214	1.37%
NOUN ADP NOUN	187	1.19%
VERB NOUN	186	1.19%
ADP ADJ	174	1.11%
DET SCONJ	174	1.11%
ADV SCONJ	171	1.09%
ADP NOUN ADP	168	1.07%
ADP ADP	165	1.05%
Other	5,658	35.98%

Table 3: Distribution of MWE annotations based on their UD part-of-speech structure.

Single-Annotation Candidates by POS- structure

- **NOUN NOUN, NOUN ADP NOUN**
- terminological candidates
 - *omejevalnik vrtljajev* ‘rev limiter’
- typical collocations
 - *kraj zločina*, lit. ‘place of the crime’, ‘scene of the crime’
- titles or functions
 - *poveljnik straže* ‘captain of the guard’

Struct.	Sin.	% (Sin.)	% (All)	Ratio
NOUN NOUN	195	3.88%	2.5%	1.55
ADP DET	172	3.42%	2.56%	1.34
PART ADV	88	1.75%	1.46%	1.2
PROPN	72	1.43%	0.63%	2.27
NOUN ADP				
NOUN	71	1.41%	1.19%	1.18
ADP PRON	68	1.35%	0.69%	1.96
VERB ADV	50	1.0%	0.54%	1.85
ADV CCONJ	46	0.92%	0.43%	2.14
SCONJ AUX	42	0.84%	0.53%	1.58
CCONJ PART	37	0.74%	0.29%	2.55

Table 4: Comparison of the distribution of part-of-speech structures between single annotations and all annotations (10 most frequent structures that are also most typical of single annotations). The columns show the number of single annotations within the structure, the percentage that structure covers within single annotations, the percentage it covers in all annotations, and the ratio between percentages.

Single-Annotation Candidates by POS-structure

- named entities or general concepts?
 - *ministrstvo za finance* ‘ministry of finance’
- partially metaphoric
 - *gostja večera*, lit. ‘guest of the evening’; ‘the guest of tonight’s show’
- sequences of prepositions and demonstrative pronouns occurring in a very vague context
 - *glede tega* ‘regarding this’, *iz tega* ‘from this’
- POS-structures with closed-class POS elements (ADP DET, ADP PRON)
 - some represent legitimate MWEs (e.g. *po svoje*, ‘in its own way’; *pri nas*, lit. ‘at us’, ‘in our country’)
 - extracting n-grams with closed-class structures

Multiple-Annotation Candidates by POS- structure

- **ADP DET NOUN** (po vsej verjetnosti ‘in all likelihood’)
- **ADP NOUN ADP** (v skladu z, ‘in accordance with’)
- **ADP ADJ** (med drugim, ‘among other things’)
- generating a list of MWEs containing closed-class elements would be useful: ADP ADP (od – do, ‘from – to’), NUM ADP (eden od, ‘one of’), DET SCONJ (več kot, ‘more than’)

Struct.	Mul.	% (Mul.)	% (All)	Ratio
VERB ADP				
NOUN	232	3.6%	2.3%	1.57
PART AUX	216	3.35%	2.25%	1.49
ADP DET				
NOUN	169	2.62%	1.9%	1.38
ADP NOUN				
ADP	127	1.97%	1.07%	1.84
NUM ADP	115	1.79%	0.99%	1.81
ADP ADP	113	1.75%	1.05%	1.67
DET SCONJ	112	1.74%	1.11%	1.57
ADP ADJ	108	1.68%	1.11%	1.51
X X	72	1.12%	0.68%	1.65
X	66	1.03%	0.54%	1.91

Table 5: Comparison of the distribution of part-of-speech structures between multiple annotations and all annotations (10 most frequent structures that are also most typical of multiple annotations).

Annotation Overlap

- Toda **[v nasprotju] s** svojimi sorodniki sodijo kaneloni (cannello = cevka) šele slabih sto let k italijanski testeninski klasiki.
 - But **contrary to** their relatives, cannelloni (cannello = tube) have been a part of the Italian pasta classics for less than one hundred years.
-
- 8,864 annotated candidates
 - 5,023 (56.67%) single annotator; **3,841 (43.33%) multiple annotations**
 - Out of 3,841 candidates, **2,961 (77.10%) exhibited complete overlap** – meaning that all the annotators annotated the exact same elements in each case

Differing Elements

- Only 880 examples showed disagreement in overlap
- prepositions (ADP), determiners (DET), pronouns (PRON), particles (PART) and conjunctions (SCONJ, CCONJ) account for more than 40% of all differing elements

UPOS	Nr.	%
ADJ	227	16.85%
NOUN	210	15.59%
ADP	172	12.77%
VERB	163	12.10%
DET	116	8.61%
AUX	116	8.61%
PRON	73	5.42%
PART	72	5.35%
ADV	62	4.60%
CCONJ	57	4.23%
SCONJ	56	4.16%
NUM	18	1.34%
PROPN	5	0.37%

Table 6: Frequencies and percentages of parts of speech causing disagreement in MWE scope annotation.

Most Frequent Disagreement in Structures

- nested MWEs?
 - varuh človekovih pravic ‘human rights ombudsman’
 - človekove pravice ‘human rights’
 - šef obveščevalne službe ‘secret service director’
 - obveščevalna služba ‘secret service’
- optional vs. obligatory elements
 - človeške pravice ‘human rights’
 - temeljne človeške pravice ‘fundamental human rights’

Diff.	Str. Pair	Freq.
ADJ	ADJ ADJ NOUN - ADJ NOUN	208
ADJ	ADP ADJ NOUN - ADP NOUN	65
NOUN	ADJ NOUN - NOUN ADJ NOUN	79
NOUN	ADJ NOUN - NOUN ADP ADJ NOUN	23
VERB	ADP NOUN - VERB ADP NOUN	90
ADP	ADJ NOUN - ADP NOUN	62
ADP	ADJ NOUN - ADP ADJ NOUN	41
AUX	AUX VERB ADP NOUN - VERB ADP NOUN	24
AUX	AUX VERB NOUN - VERB NOUN	20
DET	ADP DET NOUN - ADP NOUN	91
PART	ADP NOUN - PART ADP NOUN	19

Overlap with Named Entities

- **334 (3.77%)** candidates with at least one token that has also been annotated as a **named entity** (only 115 were annotated by multiple annotators)
- Tokens within MWEs overlapping with NEs:
 - organizations (48%)
 - no annotation (39%)
 - miscellaneous (10%)
 - location (2%)
 - person (1%)
 - person-derivative (0.5%)
- MWEs nested within NEs: *[Ustavno sodišče] Slovenije* ‘the [Constitutional Court] of Slovenia’
- NEs within MWEs: *na sončni strani [Alp]*, lit. ‘on the sunny side of [the Alps], ‘in Slovenia’

Conclusion and Future Work

- Harmonization of annotations with UniDive guidelines
- Finalization of annotations in SUK 1.0
 - Additional annotations to cover the ~11,000 sentences from PARSEME
- Further cross-analysis with other annotation layers in SUK
 - dependency relations, semantic roles

Thank you!

Jaka Čibej
jaka.cibej@ff.uni-lj.si

Polona Gantar
apolonija.gantar@ff.uni-lj.si

Mija Bon
mija.bon@ff.uni-lj.si

