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

Jaka Čibej, Polona Gantar, Mija Bon
Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, 1000 Ljubljana, Slovenia
{jaka.cibej, apolonija.gantar, mija.bon}@ff.uni-lj.si

Abstract

Recent progress within the UniDive COST Action on the compilation of universal guidelines for the annotation of non-verbal multiword expressions (MWEs) has provided an opportunity to improve and expand the work previously done within the PARSEME COST Action on the annotation of verbal multiword expressions in the SUK 1.0 Training Corpus of Slovene. A segment of the training corpus had already been annotated with verbal MWEs during PARSEME. As a follow-up and part of the New Grammar of Modern Standard Slovene (NSSSS) project, the same segment was annotated with non-verbal MWEs, resulting in approximately $6,500$ sentences annotated by at least three annotators (described in Gantar et al., 2019). Since then, the entire SUK 1.0 was also manually annotated with UD-part-of-speech tags. In the paper, we present an analysis of the MWE annotations exported from the corpus along with their part-of-speech structures through the lens of Universal Dependencies. We discuss the usefulness of the data in terms of potential insight for the further compilation and fine-tuning of guidelines particularly for non-verbal MWEs, and conclude with our plans for future work.

Keywords: multiword expressions, Universal Dependencies, Slovene

1. Introduction

Slovene was one of the languages involved in the PARSEME COST Action [1]. As part of the activities, $11,411$ sentences (approx. 41 %) of the ssj500k 2.1 Slovene Training Corpus (Krek et al., 2018)[2] were annotated with verbal MWEs (Gantar et al., 2017) categorized according to the PARSEME guidelines and MWE-tests (Savary et al., 2018). Work on Slovene MWEs within the same corpus then continued after the conclusion of PARSEME within the national project titled *New Grammar of Contemporary Standard Slovene: Sources and Methods*[3], during which non-verbal MWE annotations were added to $6,500$ sentences (a subset of the $11,411$ sentences annotated within PARSEME). Non-verbal MWEs were annotated (the process is described in more detail in (Gantar et al., 2019)) according to a set of guidelines designed primarily from the point of view of inclusion of MWEs in dictionaries, while the categorization principles followed the definitions used in the compilation of Slovene Lexical Database (Gantar and Krek, 2011) and the Digital Dictionary Database of Slovene (Kosem et al., 2021). However, the annotations have so far not been included in the SUK 1.0 corpus, pending an additional curation and resolution of crucial questions, mainly which of the annotated spans should be considered MWEs, particularly with regard to multiword combinations with varying levels of terminologicalness.

Recent advances within the UniDive COST Action[4], which among its tasks (specifically in Task 1.2) also includes the extension of the PARSEME verbal MWE annotation guidelines[5] with non-verbal MWEs, have provided an opportunity to continue the work already done on Slovene MWE annotations in the SUK 1.0 corpus within other projects, as well as to compare our own MWE-categorization with the one adopted within UniDive. At the time of writing this paper, the UniDive non-verbal MWE annotation guidelines contain no examples of Slovene MWEs, and a discussion is still underway. In addition to these examples, the lessons from the annotation of the SUK 1.0 corpus may provide a number of valuable insights during the initial phase of uni-

[1] *Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing*, IC1207 COST Action, 2013-2017: https://typo.uni-konstanz.de/parseme/

[2] Since then, the ssj500k training corpus was extended with several other datasets and underwent a rebranding, now being called the SUK 1.0 Training Corpus of Slovene (Arhar Holdt et al., 2022). In this paper, we refer to it using the new name unless we specifically refer to an older version. The SUK 1.0 corpus consists mostly of newspaper texts, magazines, and internet texts, with a small percentage of fiction and non-fiction.

[3] New Grammar of Contemporary Standard Slovene - project website: https://slovnica.ijs.si/?lang=en

[4] *Universality, Diversity and Idiosyncrasy in Language Technology*, CA21167 COST Action, 2022-2026: https://unidive.lisn.upsaclay.fr/

[5] PARSEME Annotation guidelines 1.3 - https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 154–162. May 25, 2024.

fying the PARSEME annotation scheme with Universal Dependencies (Savary et al., 2023). While the data only covers Slovene, its advantage is that several statistical calculations were made based on the annotations, including for example the scope of MWE annotation and length overlap, as well as inter-annotator agreement (each sentence was annotated by at least three annotators). In the paper, we discuss the annotated MWEs and focus predominantly on the points of disagreement and lessons learned that may prove useful for the compilation of MWE annotation guidelines within UniDive. The paper is structured as follows: we first provide a short overview of related work on MWEs (Section 2) and describe the data on annotated MWEs exported from the SUK 1.0 corpus (Section 3), then provide an analysis (Section 4). We conclude the paper with a discussion on the usefulness of the data within UniDive and a list of potential future steps in our work.

2. Related Work

MWEs still pose a problem for NLP tools such as machine translation systems, word sense disambiguation, or computational lexicography (e.g. MWE detection in corpora). A number of endeavors have been undertaken to provide training or evaluation datasets annotated with MWEs, both monolingual (Adalı et al., 2016 for Turkish; Candito et al., 2020 for French; Kato et al., 2018 and Schneider et al., 2014 for English; Mohamed et al., 2022 for Arabic; Souza and Freitas, 2023 for Portuguese) and multilingual (Monti et al., 2015; Han et al., 2020; Savary et al., 2018).

So far, no Slovene manually annotated corpus includes comprehensive and systematic annotations of MWEs; aside from the already mentioned PARSEME verbal MWE annotations in the ssj500k 2.1 Training Corpus (Gantar et al., 2017) which also serves as the Slovene UD Treebank, a small dataset for the automatic detection of idiomatic expressions has also been made by Škvorc et al. (2022) in order to facilitate idiomatic expression extraction using contextual embeddings. There is also the Slovene subcorpus of the ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al., 2021); however, MWEs within the corpus have not been categorized and only their spans have been annotated, while the corpus itself was primarily compiled for word sense disambiguation focused on single word units.

The first step toward extending the SUK 1.0 corpus with comprehensive MWE annotations was made by (Gantar et al., 2019) by conducting an experimental annotation campaign to identify potential MWE candidates. We discuss the results in the following sections.

3. Data Description

The annotation process and the typology used to annotate MWEs in SUK 1.0 was described in detail by (Gantar et al., 2019), so we only provide a brief overview here. The main goal of the task was to annotate non-verbal multiword expressions according to a typology that defines two main subgroups of MWEs[6]: (a) *lexical units*, which require an explanation (due to them being characterized by a certain degree of semantic non-compositionality), and (b) *lexico-grammatical units*, which are semantically relatively transparent (they complement or disambiguate the sense description of a headword (e.g. collocations) or they play a role of syntactic connectors or discourse organizers in language[7]).

Multiword lexical units are further divided into *fixed expressions* (which typically cover terminological expressions such as *črna luknja* 'black hole' in the sense of an astronomic phenomenon) and *phraseological units* (which typically express a metaphorical or pragmatic meaning, such as *princ na belem konju* lit. 'prince on a white horse'; 'knight in shining armor').

Lexico-grammatical units, on the other hand, consist of *collocations* (not included in the annotation task as they are regarded as semantically transparent (Atkins and Rundell, 2008) and can be automatically extracted from corpora using several criteria such as structure and statistical co-occurrence) and *syntactic combinations* (which have no lexical meaning, but are nevertheless relevant for lexicographic description because they act as adverbials, sentence connectors, and discourse organizers; such as *v skladu z* 'in accordance with').

The annotators were thus tasked with annotating MWEs as either phraseological units (PU), fixed expressions (FE), or syntactic combinations (SC). It should be noted that this is a parallel categorization, so the existing verbal MWEs annotated within PARSEME were also assigned additional categories according to this system. In this paper, we focus on the annotated UD POS-structures and patterns, not the categorization according to our own typology; more detailed results of the categorization were already presented in Gantar et al.,

[6]This categorization follows the organization of language data in the Digital Dictionary Database of Slovene (Kosem et al., 2021), where the main criterion to distinguish different types of MWEs depends on whether the MWE is a semantically independent or dependent unit.

[7]In retrospect, it should be mentioned that the decision to explicitly categorize discourse organizers as lexico-grammatical units caused some disagreement during annotation; if a discourse organizer (such as *v bistvu* 'in fact', 'actually') requires a semantic explanation and plays a pragmatic role in the sentence that needs to be explained in a dictionary, it should be categorized as a phraseological unit (PU), which falls under lexical units.

2019.

The annotation resulted in a total of 15,727 MWE annotations in the first 6,500 sentences of the SUK 1.0 corpus. Each sentence was annotated by at least 3 annotators (see Gantar et al., 2019), so a potential MWE-candidate within an individual sentence has up to three annotations (depending on whether the annotator identified the span as a MWE). For instance, in the sentence below, two annotators identified one MWE candidate and each provided an annotation; one annotated *v nasprotju* (lit. 'in contradiction') while the other annotated *v nasprotju s* (lit. 'in contradiction with').

sl Toda [[v nasprotju] s] svojimi sorodniki sodijo kaneloni (cannello = cevka) šele slabih sto let k italijanski testeninski klasiki.

en But contrary to their relatives, cannelloni (cannello = tube) have been a part of the Italian pasta classics for less than one hundred years.

A total of 8,864 MWE candidates were annotated in the corpus, consisting of 6,385 different potential MWEs.[8]

Since the annotations were made, a section of the SUK 1.0 corpus was also manually annotated with UD-part-of-speech tags, UD dependency relations, and named entities (see Arhar Holdt et al., 2023); this includes the 6,500 sentences annotated with both verbal and non-verbal MWEs, which enables us to export MWE annotations along with UD part-of-speech tags, dependency relations, and named entities, and observe potential patterns as well as points of potential disagreement. We provide a thorough analysis in Section 4 below.

4. Analysis

As shown in Table 1, the MWE candidates were annotated by 10 annotators; two of which (A and B) were reference annotators involved in the compilation of the annotation guidelines. The rest were students of linguistics at the University of Ljubljana. The distribution of annotations and the average number of MWE annotations per sentence shows that most of the annotators annotated MWEs similarly frequently to the reference annotators (approx. 0.5–0.6 MWEs per sentence), with two outliers, who were either too liberal (annotator I) or too strict (annotator J).

Out of 8,864 annotated MWE candidates, 5,023 (56.67%) were assigned a single annotation, 2,103 (23.73%) two annotations, and 1,738 (19.61%) three or more annotations. As shown in Table 2, a large portion of single annotations (almost 40%) were made by the most liberal annotator (I), but a significant percentage was provided by other annotators as well, including one of the reference annotators (B, with approx. 11%).[9] As the identification of MWEs is a difficult task, a certain degree of disagreement is to be expected. In the following subsections, we further analyze the annotations in order to discover any recurring misinterpretations that could point to potential gaps in the annotation guidelines.

Ann.	MWEs	Sent.	%	MWE/Sent.
A	292	500	1.86%	0.584
B	3,111	6,500	19.86%	0.479
C	1,742	2,000	11.12%	0.871
D	1,716	2,000	10.95%	0.858
E	1,124	2,000	7.17%	0.562
F	1,367	2,000	8.73%	0.683
G	903	2,000	5.76%	0.452
H	1,467	2,000	9.36%	0.734
I	3,563	2,000	22.74%	1.782
J	382	2,000	2.44%	0.191

Table 1: Table of MWE-annotations showing individual annotators, the number of MWE-annotations they made in the corpus, the number of all sentences annotated by them, the percentage of all annotations made, and the number of MWEs per sentence.

Ann.	Single cand.	%
I	1,953	38.86%
D	696	13.86%
C	601	11.96%
B	547	10.89%
H	380	7.57%
F	313	6.23%
G	307	6.11%
E	126	2.51%
J	91	1.81%
A	9	0.18%

Table 2: Distribution of single-annotation MWE candidates across annotators.

4.1. Part-of-Speech Structure

Based on the annotated tokens and their UD part-of-speech tags, the annotated MWE candidates cover 920 different structures, with the top 17 accounting for approx. 65% of all annotations (see Table 3). Each of these covers more than 1% of the annotations, while the other categories cover less. The majority of the annotations are non-verbal, with verbs

[8] The 6,385 different candidates were counted based on the alphabetical combinations of lemmas within annotated spans.

[9] A more detailed calculation of MWEs missed or intentionally left unannotated by individual annotators can be made once the final annotations have been encoded in the corpus.

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.

featured in only two of the most frequent categories (VERB ADP NOUN and VERB NOUN). The most frequent part-of-speech structure is ADJ NOUN (e.g. *sodni postopek*, 'judicial process', *vozniško dovoljenje*, 'driver's license'), covering almost a third of all annotations, and ADP NOUN (e.g. *v celoti*, lit. 'in whole', 'entirely'; *pred časom*, lit. 'before time', 'some time ago').

We analyzed the distribution of the part-of-speech structures in terms of how prone they were to single annotations in order to check whether any structure is more problematic for MWE identification. Table 4 shows the 10 most frequent part-of-speech structures that are also more typical of single annotations compared to all annotations (i.e. according to the ratio in the last column, they are more likely to be annotated by just a single annotator and less likely to be annotated multiple times).

An analysis of the single annotation examples with these structures reveals a number of problematic groups, particularly within structures with a nominal distribution (e.g. NOUN NOUN, NOUN ADP NOUN). First, there are terminological candidates that may be somewhat compositional, but have a specific meaning within a certain field (e.g. *omejevalnik vrtljajev* 'rev limiter', *raziskave tržišča* 'market research', *vitamin C*, 'vitamin C'). In some cases, the annotated spans are collocations that are semantically transparent, but very typical (e.g. *kraj zločina*, lit. 'place of the crime', 'scene of the crime'; *balzam za ustnice*, 'lip balm'). Secondly, some spans denote titles or functions (e.g. *poveljnik straže*, 'captain of the guard'; *hranilec družine*, lit. 'feeder of the family', 'family provider') or even members of an association or organization (e.g. *sestre usmiljenke*, 'Sisters of Mercy'), which should be treated more as named entities despite not being capitalized. Similarly, the third problematic group contains spans that can be interpreted as named entities, but that is not entirely clear when the span is spelled with no capitalization and the context is somewhat ambiguous whether the examples refer to concrete instances or a general concept (e.g. *liga prvakov*, 'league of champions'; *ministrstvo za finance*, 'ministry of finance'). In addition, examples contain phrases in which one of the components exhibits a metaphoric meaning - e.g. *gostja večera*, 'guest of the evening' in the sense of 'the guest of tonight's show'), which prompts the annotator to treat the span as non-compositional.

Next, there are several grammatical constructions that were mistakenly annotated as multiword expressions, such as combinations of prepositions and relative pronouns (ADP DET; *v kateri* 'in which', *po kateri* 'after which', *h kateri* 'to which'); some of the annotators probably annotated these because *kateri* as a relative pronoun only occurs next to prepositions, so they treated both components as a single unit. Similarly, sequences of prepositions and demonstrative pronouns (*glede tega* 'regarding this', *iz tega* 'from this') occurring in a very vague context could have prompted to treat them as non-compositional, as in the example below:

sl Država **s tem** priznava, da je prostovoljnih vojakov premalo, če ne kar nič.

en **With this**, the State recognizes that there are too few voluntary soldiers, if any.

Interestingly, some candidates with similar part-of-speech structures (either ADP DET or ADP PRON) do represent legitimate MWEs (e.g. *po svoje*, 'in its own way'; *pri nas*, lit. 'at us', 'in our country'), but were only annotated once, which indicates that expressions containing mostly closed-class parts-of-speech (which frequently constitute syntactic combinations according to our typology) should be described in more detail in the guidelines, with additional negative examples. Before manually annotating additional sentences in the corpus, a more targeted approach could be taken by extracting n-grams with problematic closed-class structures and creating a list of all syntactic combinations discovered this way (e.g. two-part connectors such as *ne samo A, temveč tudi B* 'not only A, but also B').

Table 5, on the other hand, shows the part-of-speech structures that were more likely annotated by multiple annotators (3 or more). The most frequent structure, VERB ADP NOUN (e.g. *vzeti pod drobnogled*, lit. 'take [sth] under the microscope', 'to take under scrutiny'), was frequently and consistently annotated because it contains verbal MWEs previously annotated with PARSEME categories

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.

(which the annotators followed). The second structure (PART AUX) contains just one MWE, *naj bi*, which is a very crystallized expression used in the sense of 'is said to', and was mentioned in the guidelines as a good example of a syntactic combination. Among the more intuitive structures are ADP DET NOUN (*po vsej verjetnosti*, 'in all likelihood'; *do te mere*, 'to such a degree'), ADP NOUN ADP (*v zvezi z*, lit. 'in connection with', 'with regard to', *v skladu z*, 'in accordance with'), and ADP ADJ (*med drugim*, 'among other things'; *pred kratkim*, 'a short while ago'). Some structures confirm that generating a list of MWEs containing closed-class elements would be useful: for instance, ADP ADP (*od - do*, 'from - to'), NUM ADP (*eden od*, 'one of') and DET SCONJ (*več kot*, 'more than') were quite consistently annotated because they were listed in the guidelines. The same goes for abbreviations (X and X X, such as *itn.*, *in tako naprej*, 'and so on'; *t. i.*, *tako imenovani*, 'so-called'), which could also be extracted and included in a reference list.

The two most frequently annotated structures in general (ADJ NOUN and ADP NOUN) appear almost equally frequently in both the single annotations as well as multiple annotations. This is to be expected, as the difference between a MWE and, for instance, a collocation or a terminological candidate, is a question of semantic interpretation, particularly in the context of the guidelines used for this annotation task, which relied heavily on the annotator's interpretation on whether an annotated span would require a semantic or encyclopedic explanation in a (general) dictionary language resource.

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).

4.2. Annotation Scope and Overlap

In this section, we analyze the degree to which the annotators agreed on the scope of the annotation of individual MWE candidates. Out of the $8,864$ annotated candidates, $5,023$ (56.67%) were annotated by a single annotator, while $3,841$ (43.33%) were assigned multiple annotations. Out of these $3,841$ candidates, $2,961$ (77.10%) exhibited complete overlap, meaning that all the annotators annotated the exact same elements in each case. The vast discrepancy between single annotations and the percentage of candidates with complete overlap indicates that while there is disagreement on whether a span is a MWE, in the majority of examples where a span is identified as a MWE by multiple annotators, they tend to agree on the elements included. Only 880 examples showed disagreement in annotation scope. For each candidate with incomplete overlap, we first aggregated all the annotated elements and identified the ones that differed between the annotations. In the example below, the MWE candidate was independently annotated four times (*Prav tako*, *tako kakor*, *Prav tako*, *Prav tako kakor*). Only the element *tako* (ADV) appears in all annotations, while *prav* (PART) and *kakor* (SCONJ) do not, so they are treated as differing elements.

sl **Prav tako** jasen **kakor** prejšnji, bilo je le nekoliko hladneje.

en **Just as** clear **as** the day before; it was only somewhat colder.

Table 6 shows the distribution of differing elements by part-of-speech. While adjectives and nouns are at the top of the list, prepositions (ADP), determiners (DET), pronouns (PRON), particles

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.

(PART) and conjunctions (SCONJ, CCONJ) account for more than 40% of all differing elements.

To identify potential recurring points of disagreement within specific part-of-speech structures, we also exported co-occurrences of differing structures from annotations with incomplete overlap. So for the example above (*prav tako kakor*), all the different structures were the following: *Prav tako*, PART ADV; *tako kakor*, ADV SCONJ, *Prav tako*, PART ADV; *Prav tako kakor*, PART ADV SCONJ. We counted all the possible combinations of two (excluding the ones with equal pairs) to obtain counts of the most frequently co-occurring structures. 4,063 co-occurrences of differing structure pairs were counted and further analyzed; a selection of the most interesting pairs is shown in Table 7.

The examples in which an adjective was the contested element reveal some interesting insights: the ADJ ADJ NOUN - ADJ NOUN dilemma raises the issue of annotating potential nested MWEs (*varuh človekovih pravic*, 'human rights ombudsman' vs. *človekove pravice*, 'human rights'), as well as the issue of optional vs. obligatory elements in MWEs (e.g. *človeške pravice*, 'human rights', vs. *temeljne človeške pravice*, 'fundamental human rights'). This is similar to ADP ADJ NOUN - ADP NOUN (*po ocenah*, 'according to estimates' vs. *po prvih ocenah*, 'according to the first estimates'). While the guidelines provided instructions on how to treat some of the optional elements, they were mainly focused on the inclusion of verbs in examples such as *pisati na roko*, 'to write by hand'). As a general rule, however, each example was to be annotated individually based on how typical the syntactic environment of the identified MWE was, along with the relevant lexical elements. For further annotation, the treatment of these elements should be further specified in order to avoid disagreement.

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
CCONJ	ADP DET - ADP DET CCONJ	35
NUM	ADP NOUN - ADP NUM NOUN	15

Table 7: Most frequent co-occurring structures within annotations with incomplete overlap. The first column denotes the differing element, the second the structure pair, and the third the frequency of co-occurrence.

When nouns are the differing element, the examples again show some discrepancy when it comes to potential nested MWEs (e.g. ADJ NOUN - NOUN ADJ NOUN; *ponudniki mobilnih signalov*, 'mobile signal providers' vs. mobilni signal, 'mobile signal'; *šef obveščevalne službe*, 'secret service director' vs. *obveščevalna služba*, 'secret service'; or ADJ NOUN - NOUN ADP ADJ NOUN; *rak na maternicnem vratu*, lit. 'cancer on the uteral neck', 'cervical cancer' vs. *maternični vrat*, 'cervix'). The current annotation task did not include the annotation of nested MWEs, but the results show that the guidelines should be extended to address this topic and provide clearer instructions (either by allowing for nested annotations or by listing principles on how to determine the optimal scope of the MWE).

The examples with verbs as the differing element seem to indicate that the pool of available lexical candidates that can be substituted within a MWE affects the annotator's scope. For instance, the structure pair ADP NOUN - VERB ADP NOUN contains both the verbless *na voljo*, 'at [someone's] disposal' as well as *imeti na voljo*, 'to have at [someone's] disposal', *dati na voljo*, 'to put at [someone's] dis-

posal', *biti na voljo*, 'to be at [someone's] disposal'. The relatively low number of verbs that can be used with *na voljo* seemed to prompt most, but not all of the annotators to include the verb, while others left it out.

Prepositions were frequently contested when in combination with a nominal phrase, e.g. ADJ NOUN - ADP NOUN (*v smislu* 'in [the] sense' vs. *formalnem smislu*, 'formal sense'; *v letih*, 'in [the] years' vs. *zadnjih letih*, 'last years') or ADJ NOUN - ADP ADJ NOUN (*[na] delovnem mestu* '[in] the workplace', *[v] zrelih letih*, lit. 'in mature years', 'at an older age'). Annotators were instructed to consult Slovene corpora to determine the most frequent scope of annotation, but while some interpreted the preposition as an obligatory element, others left it out based on their interpretation, e.g. whether the adjective in the MWE can be considered an open slot (*v [zadnjih/prejšnjih/naslednjih] letih*, 'in the [last/previous/next] years'; similar to numerals in ADP NOUN - ADP NUM NOUN: *pred [desetimi] leti*, '[ten] years ago'; or determiners in ADP DET NOUN - ADP NOUN: *čez [nekaj] dni*, 'in [a few] days') or whether the nominal phrase occurs frequently enough by itself (*delovno mesto*, 'workplace').

There is also some disagreement with regard to the inclusion of auxiliary verbs in verbal MWEs, e.g. AUX VERB ADP NOUN - VERB ADP NOUN (*[je] vzel pod drobnogled*, '[did] take under scrutiny') and AUX VERB NOUN - VERB NOUN (*[ni] odprl ust*, lit. '[didn't] open [his] mouth', 'remained silent'), particularly when there is a negation, but both the negated and non-negated versions are viable (*je odprl usta*, 'he spoke', *ni odprl ust*, 'he remained silent').

4.3. Overlap with Named Entities

Because the SUK 1.0 corpus was also independently annotated with named entities, we analyzed our MWE annotations in terms of tokens that have been annotated as named entities in order to explore any potential legitimate overlaps. Only 334 (3.77%) candidates contain at least one token that has also been annotated as a named entity, and only 115 were annotated by multiple annotators. By analyzing the distribution of the named entity annotations within these 115 candidates, we see that the majority were annotated as organizations (48%) or have no annotation (39%; meaning that not all the MWE elements overlap with the named entity), while other NE categories account for much smaller percentages: miscellaneous (10%), location (2%), person (1%), and person-derivative (0.5%). The guidelines mention that generic titles of institutions, documents, etc. should be annotated as MWEs, particularly if they indicate culturally specific expressions with no direct equivalents or transparent translations in other languages.

A closer look at the examples shows that in the majority of cases, the MWE annotations are nested within NE annotations (e.g. *[Ustavno sodišče] Slovenije*, 'the [Constitutional Court] of Slovenia'; *Urad za [narodnostne manjšine]*, 'Office of [National Minorities]'), but the opposite also occurs, with NEs included in MWEs (*na sončni strani [Alp]*, lit. 'on the sunny side of [the Alps], 'in Slovenia'; *kdor gre na [Dunaj], naj pusti trebuh zunaj*, lit. 'whoever goes to Vienna should leave their stomach outside', 'Vienna is very expensive' or 'large cities are very expensive') or appearing in open slots of MWEs (*so voda na [Lutov] mlin*, lit. 'they are water to [Lut's] mill', 'they provide an advantage to him'). These examples are useful to include in the improved guidelines to exemplify the interplay between MWEs and NEs and to provide clearer instructions on how to annotate mixed candidates.

5. Conclusion

In the paper, we presented the results of the first step of the process of comprehensive MWE annotation in the SUK 1.0 corpus, and conducted a number of quantitative analyses to pinpoint potential weak points in the first version of our annotation guidelines. In particular, the process shows that more instructions and examples are required on how to differentiate between terminological candidates and collocations on one hand, and MWEs on the other. Although the annotators seem to achieve a considerable degree of overlap in terms of annotation scope, for some structures, the scope should be more precisely defined (e.g. the inclusion of auxiliary verbs and closed-class parts-of-speech such as prepositions). In addition, closed-class part-of-speech structures can be pre-extracted in order to generate a list of valid candidates as a reference point for annotators and, potentially, for pre-annotating some of the more trivial syntactic combinations. Pre-annotation with a list of all other MWE-candidates is also an option, but might be more difficult to implement for Slovene, which features a flexible word order and is a morphologically rich language.

Although there has not been much overlap between MWEs and NEs in the annotated examples, the ones that do occur nevertheless show the need for more specific guidelines on when to treat candidates as named entities and how to treat borderline examples (e.g. when the lack of capitalization makes it unclear whether the span denotes a named entity or a generic concept) and mixed candidates (nested MWEs within NEs or vice versa).

In our future work, we intend to use the UniDive MWE annotation guidelines to perform a second step annotation of the identified MWE candidates

and determine their categories so that they can be added to the SUK 1.0 corpus alongside their PARSEME verbal MWE equivalents. Once the final annotations have been added to the corpus, a second analysis of outlying examples (either those left unannotated by the majority of annotators or those consistently annotated but not considered MWEs in the final version) can provide additional insight for further MWE identification. In addition, the annotated POS-structures can potentially be compared to the total frequencies of POS-structures within the corpus in order to pinpoint whether certain structures are more typical of MWEs in Slovene in general. Additional statistical analyses on MWE patterns can also be performed by taking into account other annotation layers present in the corpus, such as semantic role labeling and UD dependency relations.

6. Acknowledgements

The study presented in this paper was conducted within the *New Grammar of Modern Standard Slovene: Resource and Methods project* (J6-8256), which was financially supported by the Slovenian Research and Innovation Agency (ARIS) between 2017 and 2020. The authors also acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 - *Language Resources and Technologies for Slovene* and No. P6-0215 - *Slovene Language – Basic, Contrastive, and Applied Studies*).

The authors would like to thank the anonymous reviewers for their valuable insight, and all the annotators who participated in the project: Anna Maria Grego, Tjaša Šoltes, Tajda Liplin Šerbetar, Pia Rednak, Jana Vaupotič, Zala Vidic, Karolina Zgaga, and Kaja Gantar.

7. Ethical Considerations and Limitations

It should be noted that 80% of the people who performed the annotation were university-level students of linguistics, and while they were familiarized with the guidelines and their performance was tested and compared to the performance of experts and considered to be satisfactory in the majority of cases, the annotations need to be interpreted with their background in mind.

In addition, the SUK 1.0 corpus mostly contains written standard Slovene, so the results cannot necessarily be extrapolated to e.g. spoken or non-standard Slovene.

8. Bibliographical References

Kubra Adalı, Tutkum Dinc, Memduh Gokırmak, and Gülşen Eryiğit. 2016. Comprehensive annotation of multiword expressions in turkish. *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*, pages 60–66.

Špela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2022. Training corpus SUK 1.0. Slovenian language resource repository CLARIN.SI.

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Simon Krek, Tina Munda, Nejc Robida, Luka Terčon, and Slavko Žitnik. 2023. Nadgradnja učnega korpusa ssj550k v suk 1.0. *Razvoj slovenščine v digitalnem okolju*, pages 119–156.

B. T. Sue Atkins and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford Univer-sity Press, New York.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2020. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*.

Polona Gantar, Jaka Čibej, and Mija Bon. 2019. Slovene multi-word units: Identification, categorization, and representation. In *Computational and Corpus-Based Phraseology*, pages 99–112, Cham. Springer International Publishing.

Polona Gantar and Simon Krek. 2011. Slovene lexical database. *Natural Language Processing, Multilinguality: Sixth International Conference*, pages 1–13.

Polona Gantar, Simon Krek, and Taja Kuzman. 2017. Verbal multiword expressions in slovene. *International Conference on Computational and Corpus-Based Phraseology*, pages 1–13.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Iztok Kosem, Simon Krek, and Polona Gantar. 2021. Semantic data should no longer exist in isolation: the digital dictionary database of slovenian. *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion*, pages 81–83.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2018. Training corpus ssj500k 2.1. Slovenian language resource repository CLARIN.SI.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña Ruiz, José Luis Sancho Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, and Tina Munda. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, pages 377–395.

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskandar Keskes, Jean-Yves Antoine, and Lamia Belguith Hadrich. 2022. Annotating verbal multiword expressions in arabic: Assessing the validity of a multilingual annotation procedure. *13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1839–1848.

Johanna Monti, Federico Sangati, and Mihael Arčan. 2015. Ted-mwe: a bilingual parallel corpus with mwe annotation: Towards a methodology for annotating mwes in parallel multilingual corpora. *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 193–197.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati Sangati, Ivelina Stoyanova Stoyanova, and Veronika Vincze. 2018. Parseme multilingual corpus of verbal multiword expressions. *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. Parseme meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tadej Škvorc, Polona Gantar, and Marko Robnik Šikonja. 2022. Mice: mining idioms with contextual embeddings. *Knowledge-based systems Jan. 2022, vol. 235*, pages 1–11.

Elvis Souza and Claudia Freitas. 2023. Annotation of fixed multiword expressions (MWEs) in a Portuguese Universal Dependencies (UD) treebank: Gathering candidates from three different sources. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 442–450, Belo Horizonte, Brazil. Association for Computational Linguistics.