Polona Gantar/Simon Krek

CREATING THE LEXICON OF MULTI-WORD EXPRESSIONS FOR SLOVENE

Methodology and structure

Abstract This paper describes a method for automatic identification of sentences in the Gigafida corpus containing multi-word expressions (MWEs) from the list of 5,242 phraseological units, which was developed on the basis of several existing open-access lexical resources for Slovene. The method is based on a definition of MWEs, which includes information on two levels of corpus annotation: syntax (dependency parsing) and morphology (POS tagging), together with some additional statistical parameters. The resulting lexicon contains 12,358 sentences containing MWEs extracted from the corpus. The extracted sentences were analysed from the lexicographic point of view with the aim of establishing canonical forms of MWEs and semantic relations between them in terms of variation, synonymy, and antonymy.

Keywords Identification of multi-word expressions; multi-word expressions lexicon; canonical form of multi-word expressions; Slovene; Digital Dictionary Database

1. Introduction

Multi-word units (MWEs) require a number of decisions to be made in relation to their linguistic and communicative properties both from a lexicographic and a natural language processing perspective. From a lexicographic point of view, the main issue concerning MWEs is understanding their type and using appropriate methods for their lexicographic description. For instance, lexicographers need to decide if all types of MWEs require explicit semantic interpretation in the form of an explanation or definition. In other words, they need to establish how distant the meaning of a phrase as a whole is from the meanings of its individual components. Another typical lexicographic issue is the placement of MWEs in the macro- and microstructure of a dictionary. Mistakenly, this question is thought to have been made obsolete by the decline of paper dictionaries. Rendered in a different manner, it remains relevant in e-dictionaries or in Digital Dictionary Databases. In born-digital dictionaries (cf. Tavast et al. 2018; Tiberius et al. 2021; Gantar 2020), there is a trend to include different types of MWEs as stand-alone lexical units or headwords. However, other modern dictionaries tend to incorporate MWEs under their individual components, which function as dictionary headwords; this raises the issue of establishing semantic links between the MWE as a whole and the senses of its individual components. Lastly, there are also issues related to the canonical form of a MWE: which of its possible variations represents the canonical form, what is the canonical word order, the number of components, relations between components, and in particular, how do we treat the numerous variations and syntactic transformations between semantically similar MWEs.

On the other hand, natural language processing is mainly concerned with two MWE-related tasks: the identification of MWEs in a running text, based on an existing MWE lexicon, and the extraction of MWEs from a text for lexicon building (Savary/Cordeiro/Ramisch 2019). For the first task, either a list of MWEs from existing lexicons or manually annotated corpora are used as the basis (Ramisch et al. 2020), and the result is a set of corpus sentences con-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

taining MWEs in all typical syntactic and semantic realisations. The second procedure refers to the identification of MWEs in corpora regardless of existing MWEs.

For our purposes, the first approach is more appropriate because it identifies both MWEs appearing in the expected or pre-recorded syntactic structures and their lexical realisations, as well as unregistered word combinations that are potentially lexicographically relevant. Our goal is to build a lexicon of automatically extracted MWEs and to integrate all the extracted data into the Digital Dictionary Database.

In this paper we first describe a method for identification of MWEs in the Gigafida corpus. We focus on the extraction of MWEs with phraseological properties that are characterised by semantic opacity and metaphoricity (hereafter, the term MWE will be used in this sense). The method is based on identifying MWEs in the corpus by using various types of information about individual MWEs, e. g. their syntactic structure, underlying POS information, lemmas, etc. In particular, we highlight the use of syntactic information. First, we describe the construction of an initial list of MWEs, where we specify the obligatory and optional constituents, the expected valency slots, and the order of the constituents. We then describe the process of extracting the MWEs from the corpus and the result: a MWE lexicon, which – in its first version – contains 5,242 MWE lexical units, together with morpho-syntactic information on their components, and 12,358 examples from the Gigafida 2.0 corpus containing corresponding MWEs.

In the second part of the paper, we describe lexicographic aspects of using this approach – the analysis of the extracted data. Our starting point is that the semantic value of MWEs needs to be considered in order for us to be able to establish the links between their numerous variants. Therefore, we will be interested in how MWEs with overlapping constituents and potentially similar meaning are related to each other, considering their variance and syntactic transformations. The lexicographic analysis also aims to identify linguistic characteristics which have a predominant influence on the meaning of MWEs and use this to establish semantic links (variation, synonymy, antonymy) that we want to include in the Digital Dictionary Database.

2. Methodology of MWE identification in the corpus

Our method for identifying MWEs in text corpora is based on a formal description of syntactic structures encoded in the XML format. The previous method we used for this purpose was based on the formal description of grammatical relations in the Word Sketches tool for Slovene (Krek/Kilgarriff 2006), i. e. the sketch grammar. The upgraded method also takes into account the level of syntactic annotation according to the dependency model, in addition to the underlying morphological level of annotation and some further statistical parameters. This method enables us to control the morphological properties of the MWE components as well as the syntactic relations between the components and provides the opportunity to work with both lexical and syntactic variations, which profoundly influence the choice of the canonical form of the MWE in the Digital Dictionary Database.

2.1 The Initial list of MWEs

To create the initial list of MWEs, we used three lexical resources which are freely available and include MWE types that match the required properties: Slovene Lexical Database (Gantar et al. 2013), the Dictionary of Slovene Phrasemes (DSP) (Keber 2011), and a list of idioms that were manually tagged in the ssj500k 2.0 training corpus (Krek et al. 2020) based on guidelines set out in the PARSEME COST Action (Bhatia et al. 2017). The final list, which was edited according to the criteria described below, comprised 5,241 MWEs.

Typologically, we relied on the Digital Dictionary Database (DDD) model for the selection of MWEs that would represent the initial set. DDD includes several types of MWEs (Kosem/Krek/Gantar 2020): fixed phrases – semantically independent units or terminological units with a defined syntactic structure (e.g. *okrogla miza* = round table (event), *sir s plemenito plesnijo* = blue cheese); phraseological units – semantically independent units with a primary connotative, metaphorical and/or pragmatic role (e.g. *dati mir* – *give peace 'not to disturb'); collocations – statistically relevant phrases in predictable syntactic structures that have no independent semantic value as a whole (e.g. *angleški, slovenski, italijanski … jezik* = English, Slovenian, Italian … language); syntactic combinations – semantically transparent or non-transparent phrases with a predictable syntactic structure and sentence role, e.g. *na podlagi česa* = on the basis of something; *pod okriljem koga* = under the auspices of somebody. Among these types only phraseological units were taken into account in the creation of the first version of the lexicon. The list was manually edited by removing duplicate MWEs appearing either within a single source (for example, each MWE appears in the DSP as many times as the number of components it contains) or in more than one source. The list was re-arranged by identifying the number of components for each MWE, including optional lexical components (indicated by the bracketed part in the following example): *pasti komu v naročje (kot zrela hruška)* = *to fall in someone's lap (like a ripe pear) 'accidentally acquire something by luck', and components indicating abstract valency slots expressed by pronouns (*zlesti komu pod kožo* = to get under someone's skin 'to be able to understand or affect someone'). We have kept variants of MWEs on the list, e.g. *dati/dobiti zeleno luč* = to give/get green light, which are – as expected – syntactically and semantically related to each other, e.g. *začaran krog* = vicious circle <==> *vrteti/znajti se v začaranem krogu* = to be/to spin in a vicious circle <==> *pasti v začaran krog* = to fall into a vicious circle <==> *izvleči se iz začaranega kroga* = to get out of a vicious circle. We have removed other types of MWEs from the list, e.g. fixed phrases (črna skrinjica = black box) and collocations (*low price, low temperature*).

2.2 Syntactic structures

For automatic identification, the MWEs on the initial list were parsed and assigned with syntactic structures that form the basis for the extraction of corpus sentences. The formal description of a syntactic structure contains different types of information: (1) syntactic parse according to the dependency model, morphological information – POS + linguistic features (case, number etc.), and some additional statistical parameters. It is the dependency layer in particular which represents an upgrade from the sketch grammar model in the MWE extraction from the corpus; it is used for defining the scope (lexicalised, non-lexicalised elements) and internal structure (word forms, dependencies) of a MWE. In this section we describe the concept of a syntactic structure used in our formalism and its description in the XML format.

Each MWE is assigned with its syntactic structure or, more precisely, MWEs are grouped according to the syntactic structure they belong to. MWEs with the same (syntactic, morphological, etc.) characteristics are thus assigned with the same syntactic structure. A syntactic structure is defined by the <syntactic_structure> element, which requires three obligatory attributes: the structure ID – @id, human-readable code – @label and the structure type – @type. An example for the MWE *kaj ne da miru komu* 'something is bothering someone':

(1) <syntactic_structure type="other" label="z-l-gg-s2-z" id="122">

There are three types of syntactic structures: »single« (11 structures for each POS of the single words), »collocation« (82 structures), and »other« for all the other structures which do not belong to the first two. The syntactic structure is also attributed with a human-readable code, which contains a formalised combination of codes or tags used for POS corpus annotation, for each of the components, e. g. the syntactic structure:

(2) label="z-l-gg-s2-z"

for the mentioned MWE (*kaj ne da miru komu*) can be interpreted as a sequence of five components, with the first one being a pronoun (z-*kaj*), the second a particle (l-*ne*), the third a main verb (gg-*dati*), the fourth a noun in the genitive (s2-*miru*), and the fifth again a pronoun (z-*komu*). As already mentioned, MWEs with the same structure have the same IDs. For example, *začarani krog* (vicious circle) and *bela zastava* (white flag) both receive the same syntactic structure ID: p0-s0 (adjective-noun combination in all grammatical cases).

Tags described in the previous paragraph are used in the JOS Multext-East annotation system, which is also a required piece of information in the syntactic structure definition. This provides some flexibility regarding the use of different annotation systems, e. g. Universal Dependencies. Furthermore, individual syntactic structures are defined by three groups of information: (1) individual words or elements that form the MWE – »components«; (2) syntactic relations between the components – »dependencies«; and (3) possible restrictions and other information which is needed to extract the MWE – »definition«:

(3)
```
<system type="JOS">
<components order="fixed"></components>
<dependencies> </dependencies>
<definition></definition>
</system>
```

The order of the components (from 1–5 in our case) can be defined as »fixed« or »variable«, depending on the decision whether the extracted MWEs should follow the order defined in the syntactic structure definition or the one which is statistically prevalent in the corpus. An example of the variable structure is a combination of an adverb and a verb, where we want to follow corpus frequency with different types of adverbs: *debelo* [Adverb] *gledati* [Verb] (*to watch thickly ('to marvel, to be surprised') vs *priti* [Verb] *jutri* [Adverb] (to come tomorrow).

Each component could be of type »core« or »other«. Core components are an integral part of the MWE, while »other« components are used to include restrictions on the level of dependency parsing constituents, e. g. if we want to exclude the possibility that one of the MWE components is connected with a preposition.

(4)
```
<components order="fixed">
<component cid="1" type="core" label="z"/>
<component cid="2" type="core" label="l"/>
<component cid="3" type="core" label="gg"/>
<component cid="4" type="core" label="s2"/>
```

```
<component cid="5" type="core" label="z"/>
</components>
```

Syntactic relations between the components are defined within the »dependency« element. Again, we use the Slovenian dependency system JOS for defining syntactic relations, which consists of three types of relations – those connecting elements within (all types of) phrases, those connecting sentence elements, and all others. For our purposes, the first and the second type are the most important ones. Phrasal relations include labels »dol« (attribute), »del« (part of predicate), »prir« (coordination); sentence element relations include labels »ena« (subject), »dve« (direct/indirect object), »tri«, and »štiri« (two types of adverbials). The visualisation of the above mentioned MWU (*kaj ne da miru komu*) in the Q-CAT corpus annotation tool (Brank 2021) shows that there is a subject-predicate (»ena«) connection between the pronoun *kaj*, which functions as the subject, and the verb *dati* (to give), and between two types of objects (»dve«) and the verb: the direct object *mir* (‘peace’) and the indirect object *komu* (‘to somebody’). The negation particle *ne* (‘not’) is connected with the phrasal relation »del« (part of predicate).

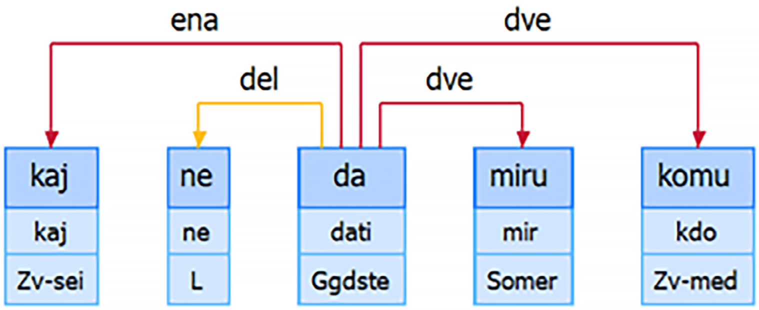

Fig. 1: A syntactically parsed MWE *kaj ne da miru komu* in the Q-Cat corpus annotation tool

The formal description of dependency relations in the XML format is the following:

(5)
```
<dependencies>
<dependency from="3" label="ena" to="1"/>
<dependency from="3" label="del" to="2"/>
<dependency from="#" label="modra" to="3"/>
<dependency from="3" label="dve" to="4"/>
<dependency from="3" label="dve" to="5"/>
</dependencies>
```

2.3 Corpus

To extract the required data, we needed a corpus annotated on various levels, based on the list of initial FEs and the definitions of syntactic structures. For this purpose, we used the Gigafida 2.1 corpus (Krek et al. 2020a), which includes additional levels of annotation, most importantly, dependency parsing annotations according to JOS (Erjavec et al. 2010, 2011) and UD (Dobrovoljc/Erjavec/Krek 2017) systems, named entities, and semantic role labelling annotations (Gantar et al. 2018).

2.4 Extraction method

Based on the input of syntactically parsed MWEs, the script identified corpus sentences with elements included in the description of each MWE. However, we wanted to go one step further and also identify possible variants or potential new MEWs. To do this, we considered the option of filling each MWE component with any other word, as shown in Table 1.

0	**barvati** **(to colour)**	**kaj** **(something)**	**s** **(with)**	**črnimi** **(black)**	**barvami** **(colours)**
	x	A	C	x	x
1	slikati (to paint)	dneve (days)	s (with)	črnimi (black)	barvami (colours)
	a	a	C	C	C
2	slikati (to paint)	nevarnosti (danger)	s (with)	črnimi (black)	barvami (colours)
	a	a	C	C	C
3	barvati (to colour)	obrazke (faces)	s (with)	pisanimi (colourful)	barvami (colours)
	C	a	C	a	C
4	barvati (to colour)	jajčka (eggs)	s (with)	posebnimi (special)	barvami (colours)
	C	a	C	a	C
5	barvati (to colour)	dogajanja (events)	s (with)	črnimi (black)	odtenki (shades)
	C	a	C	C	a
6	barvati (to colour)	kozarčke (glasses)	s (with)	posebnimi (special)	barvami (colours)
	C	a	C	a	C

Table 1: List of phraseological candidates for the MWE: *barvati kaj s črnimi barvami* (paint with black colours) extracted from the Gigafida 2.1 corpus

As shown in Table 1, when extracting sentences from the corpus, we queried the corpus for all the possible realisations (1–6) for each MWE (0) in the initial dataset. We specified the MWE components which can vary (x), the fixed components (C), and the valency slots (A). In the corpus sentences (1–6), we recorded actual lexical realisations (a) in variable components (x). These lists were then used for the semantic analysis of the real occurrences of the MWEs, as attested in the written corpus of Standard Slovene, so that we could set the rules for canonical forms in the MWE lexicon and distinguish the variants and transformations of semantically related MWEs from other independent MWEs.

3. Lexicon

Data from the MWE Lexicon is integrated into the Digital Dictionary Database as part of its data model. At the same time, it represents a specific, stand-alone resource containing 5,242 lexical units of MWE type in 12,358 examples from the Gigafida 2.0 corpus. In the Lexicon

(Example 6), a MWE (<lemma> within <headword>) is defined by a syntactic structure with its identification number. The components of the MWE are defined individually in a sequence (attribute=num) and represented in the <lexeme> element together with the lemma and morpho-syntactic description (attribute=msd). The <body> element records semantic information. Each MWE sense has its identification number (sense key) and a list of senses (RelatedSenseList) with definitions. Each MWE or its sense contains at least one and up to three examples from the Gigafida 2.1 corpus.

(6)

```
<entry>
<head>
<headword>
 <lemma>kaj ne da miru komu</lemma>
</headword>
 <lexicalUnit type=”MWE” structure_id=”122”>
<component num=”1”>
 <lexeme lemma=”kaj” msd=”Zv-sei”>kaj</lexeme>
</component>
<component num=”2”>
 <lexeme lemma=”ne” msd=”L”>ne</lexeme>
</component>
<component num=”3”>
 <lexeme lemma=”dati” msd=”Ggdste”>da</lexeme>
</component>
<component num=”4”>
 <lexeme lemma=”mir” msd=”Somer”>miru</lexeme>
</component>
<component num=”5”>
 <lexeme lemma=”kdo” msd=”Zv-med”>komu</lexeme>
</component>
</lexicalUnit>
</head>
<body>
<senseList>
<sense key=”s.24”>
<relatedSenseList>
<relatedSense senseKey=”s.26”/>
</relatedSenseList>
<definitionList>
 <definition>kaj vznemirja koga; vzbuja zanimanje pri kom</definition>
</definitionList>
<exampleContainerList>
<exampleContainer>
 <corpusExample exampleId=”GF9913201.308.2”>Hedonistična <comp
num=”1”>plat</comp> vaše osebnosti <comp num=”5”>vam</comp> <comp
num=”2”>ne</comp> bo <comp num=”3”>dala</comp> <comp num=”4”>miru</
comp>, dokler ji ne boste zares prisluhnile.</corpusExample>
</exampleContainer>
</exampleContainerList>
</sense>
</senseList>
</body>
</entry>
```

The MWE Lexicon is available in the Clarin.si repository under CC BY-SA 4.0 license.

4. Lexicographic analysis of extracted data

In section 4 we present a case study of the analysis of extracted data from a lexicographic perspective. Based on the analysis, we aim to formalise MWE descriptions in the dictionary, including the rules for canonical forms, variants, and (syntactic) transformations. All these issues are, as we will demonstrate on the example of five selected MWEs, determined by the question of their meaning. The purpose of the lexicographic analysis is thus also to identify the linguistic features which are decisive in determining the meaning of the MWEs and to incorporate these findings into a formalised description of the MWEs in the Digital Dictionary Database and, consequently, improve the system of their extraction from the corpus.

4.1 Identification of semantic features which characterise a MWE as a lexical unit

The complexity of defining a MWE as a lexical unit, which stems from, among other things, its capacity for variation and transformation, is illustrated by an example of a MWE containing two fixed components: the verb *dati* (to give) with the variant *pustiti* (to let) and the noun *mir* (peace). There were five such MWEs in the initial list: *dati mir, dati mir komu, pustiti koga pri miru, pustiti koga na miru* and *ne dati miru* (To leave alone, to leave someone alone, to leave someone in peace, to let someone be, to not let alone / to not be able to get (someone or something) out of one's mind).

In the lexicographic analysis, the extracted data were compiled in an excel file containing the MWE ID, the status of each component with respect to the identification procedure, the actual lexical realisations of each component, and the corpus sentence containing the MWE. The table below shows the extracted data for the five selected MWEs. We list only three extracted instances for each MWE, even though there are many more extracted sentences (see last column). All of the examples were then considered for their relevance in terms of number, order, and form of components, using various editing and filtering options.

comp1	**comp2**	**comp3**	**comp4**	**corpus sentence**	**freq.**
dati	mir				879
dali	mir			*Mi smo skrbeli za red, oni so dali mir.* (We maintained order, and they let us be.)	
dal	mir			*Dvakrat na dan sva hodila lulat, drugače je pa pod mizo ležal in je dal mir.* (We went out to pee twice a day, but other than that he lay under the table and left me alone.)	
dati	mir	komu			446
dajte	mir	mi		*Samo ponoči mi dajte mir!* (At least leave me alone at night!)	
dam	mir	ti		*Daj mi ga, pa ti dam mir, je rekel.* (Give it to me, and I'll leave you alone.)	

comp1	comp2	comp3	comp4	corpus sentence	freq.
ne	dati	miru			449
ne	dali	miru		*Dokler ne bomo tega dosegli, ne bomo dali miru!* (We will not rest until we achieve this.)	
ne	dal	miru		*Nekdo vam očitno še ne bo dal miru.* (Someone is obviously not going to leave you alone just yet.)	
pustiti	koga	pri	miru		3548
pustili	ga	pri	miru	*Pustili so ga pri miru.* (They left him alone.)	
pustili	vas	pri	miru	*Tudi otroci vas ne bodo pustili pri miru, zahtevali bodo vašo pozornost in ljubezen.* (Children won't leave you alone either, they will demand your attention and love.)	
pustiti	koga	na	miru		398
pustijo	te	na	miru	Če ne delaš problemov, te pustijo na miru. (If you don't cause any problems, they leave you alone.)	
pustite	ženske	na	miru	Ženske pustite na miru! (Leave the women alone!)	

Table 2: A sample of extracted phraseological candidates for 5 MWEs with components *dati/pusti* and *mir*, from the Gigafida 2.1 corpus

4.1.1 Frequency and word forms

Important initial lexicographic conclusions can be reached simply by examining the frequency data: as the extracted sentences show, the MWE *pustiti koga pri miru* (3548) has the highest number of occurrences of the five, especially in combination with the (possible) variant with the preposition *na* (398), which seems to be less frequent. We can assume that the MWE *dati mir* (879) also incorporates all the sentences for *dati mir komu* (398), which means that the same instances are extracted from the corpus for both MWEs. However, *dati mir* (879) also includes realisations without the free valency slot (*komu*), which is important as the presence or absence of this slot may affect the overall semantic interpretation of the MWE. In relation to the two MWEs under consideration, the numerical data on the representation of verb forms also provided useful information, as in the examples *dajte no mir!* and *daj mi mir!* the imperative clearly stands out among the verb forms.

4.1.2 Negation

Negation is another feature that can influence the meaning of a MWE, as well as its scope and the properties of its components. In itself, negation is a problematic category, since it can be expressed syntactically in very different ways. In our case it is expressed through the negation particle *ne*, which is detached from the verb (i. e. it is not part of the verb, as in the case of *nimam* (I do not have) or *nisem* (I am not)). The affirmative-negative pair *dati mir* and *ne dati miru* is interesting because, although seemingly opposite, the affirmative form actually denotes the opposite situation (that there is no peace), which is reflected in the imperative forms: *Dajte mir!* (Leave us alone! Let us be!). On the other hand, the MWE with the negation particle *ne dati miru* (27) does not actually signify the negated action, but its affirmation (to bother, to harass). Accordingly, this MWE is not typically used in imperative verb forms.

4.1.3 Valency slots and semantic types

Valency slots included in the MWEs play an important role in determining their meaning. Both their presence and absence are important, as well as their lexical realisations, e. g. *pustiti koga pri miru* (leave someone (=human) alone) vs. *pustiti kaj pri miru* (leave something (non-human, activity) alone – 'not to be bothered with something'). This difference can be expressed by semantic types or other mechanisms which have been used in FrameNet (Fillmore /Johnson/Petruck 2003), Corpus Pattern Analysis (Hanks 2004; Hanks/Pustejovsky 2005), and others. In future upgrades to the Lexicon, we thus intend to define the actual realisations of the valency slots in terms of semantic types, based on the Slovene ontology of semantic types for nouns SLONEST (Kosem/Pori 2021).

The semantic role of valency slots and their lexical realisations is also obvious in the case of MWEs with negation elements. Despite their surface similarity, the MWE *kaj ne da miru komu* is in no sense semantically related to the MWE *ne dati miru*, nor to *kdo ne da miru komu*, since in the first case it signifies a disturbance caused by something, whereas in the second case this is primarily a human-caused annoyance. In the case of *kdo ne da miru komu*, the only exception is to be found with female and male agents (as a semantic role), where the MWE can be understood both in the sense of arousing interest and also as harassment: *Urban, moj sodelavec z oddelka s pijačami, mi ne da miru.* (I can't get Urban, my colleague in the drinks department, out of my mind. / Urban, my colleague in the drinks department, won't leave me alone.)

In the agent position represented by the pronoun *kaj* (something), one typically finds expressions such as conscience, curiosity, thought, question, etc. Some of the words in this position, depending on their frequency and semantic nuances, could suggest independent MWEs, e. g. žilica ne da miru komu (*his/her knack (for something) won't leave him alone = knack in the sense of 'ability, talent, interest'), *hudič/vrag ne da miru komu* (*devil doesn't leave him alone), which can be interpreted as 'to do something exciting, controversial (despite everything)'.

4.2 Determining MWE scope and canonical forms

The above described characteristics provided us with a starting point for defining the canonical forms of MWEs, with all the relevant information in terms of structures and individual components, as shown in the first section. Based on the lexicographic analysis, the initial list of 5 MWEs produced 13 independent MWEs. We justified their independence in terms of their semantic properties, their scope, and the number and sequence of components. At the same time, their semantic independence is demonstrated by the specific asymmetric links between the MWEs and their senses, as shown below. We first list the semantic descriptions or identified sense and then the individual MWEs identified through lexicographic analysis (new MWEs), as well as their semantic relations, which are illustrated by the corresponding sense number.

(7) Identified senses

1	ne povzročati težav	not to cause trouble
2	ne razgrajati; biti pri miru; biti »priden«	not to carry on; to be calm; to be „goody-goody“
3	prenehati nadlegovati koga	to stop harassing someone
4	v diskurzu: izraža zahtevo, opozorilo	in discourse: expressing a request, a warning
5	v diskurzu: izraža zavrnitev	in discourse: expressing refusal
6	ne se ukvarjati s čim, ne se dotikati ipd.	not to bother with something; not to touch, etc.
7	povzročati težave	to cause trouble
8	razgrajati; ne biti pri miru; ne biti »priden«	to carry on; not to be calm; not to be „goody-goody“
9	nadlegovati koga	to harass someone
10	vzbujati zanimanje, vznemirjati	to arouse interest; to disturb
11	(kljub vsemu) narediti kaj vznemirljivega, spornega	to do something exciting, controversial (despite everything)

(8)

initial MWE	**new MWE**	**semantic relation**
dati mir	dati mir	1, 2, (7, 8)*
	daj/dajte mir!	4, 1, 2
	daj/dajte no mir!	5
dati mir komu	dati mir komu	3 (9)
	daj mi mir!	4, 3
pustiti koga na/pri miru	pustiti koga na/pri miru	3
	pustiti kaj pri miru	6
	ne pustiti koga na/pri miru	9
ne dati miru	ne dati miru	7, 8 (1, 2)
	ne dati miru komu	9 (3)
	kaj ne da miru komu	10
	žilica ne da miru komu	10
	hudič/vrag ne da miru komu	11

* () – antonymy

5. Conclusions

As we have shown in this paper, MWEs can be problematic in several respects, both in terms of the way they are included and treated in the dictionary database, as well as in terms of automatic identification in the text. Their realisations within a text display significant variability since they they can adapt to a text in numerous ways, as their individual components can be interchanged with others within a phrase and take on different word forms. Additionally, other components may be interspersed between the MWE components, while some

MWEs may also appear as free combinations, i.e. without lexical meaning. We therefore believe that MWEs, like single-word units, should be treated as lemmas when they are included in the dictionary database; furthermore, we should also determine the extent of variation, i.e. the point at which the deviation from the canonical form violates the interdependence between the form and the meaning of the MWE, which is a prerequisite for the recognition of a MWE as a lexical unit in its own right.

In this paper we analysed automatically extracted sentences containing a specific MWE to identify the linguistic instruments for recognising their semantic independence. We highlighted the importance of free valency slots and their lexical or semantic values, which can be expressed in terms of semantic types, such as human, space, process, phenomenon, etc. Pragmatic usage, evident from the predominance of specific forms of the verb constituent, such as the imperative, as well as the presence of negation forms, also emerged as an important semantic identifier. Detailed lexicographic analysis was made possible by the process of MWE extraction on the basis of predefined syntactic structures, where the (relatively stable) MWE variants and transformation options were identified on the basis of frequency and other morpho-syntactic information about individual components and dependency relations between them. This enabled us to determine canonical forms of MWEs in the Digital Database and their semantic relations.

References

Bhatia, A./Bonial, C./Candito, M./Cap, F./Cordeiro, S./Foufi, V./Gantar, P./Giouli, V./Herrero, C./Iñurrieta, U./Ionescu, M./Maldonado, A./Mititelu, V./Monti, J./Nivre, J./Onofrei, M./Ow, V./ Parra Escartín, C./Sailer, M./Ramisch, C./Ramisch, R./Rizea M./Savary, A./Schneider, N./Stonayova, I./Stymne, S./Vaidya, A./Vincze, V./Walsh, A. (2017): PARSEME shared task 1.1 annotation guidelines (last updated on November 30, 2017). http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/ (last access: 24-03-2022).

Brank, J. (2021): Q-CAT Corpus Annotation Tool 1.2. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1442 (last access: 23-03-2022).

Dobrovoljc, K./Erjavec, T./Krek, S. (2017): The Universal Dependencies Treebank for Slovenian. In: Erjavec, T./Piskorski, J./Pivovarova, L./ Šnajder, J./Steinberger, J./Yangarber, R. (eds.): Proceedings of the EACL Workshop. The 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, 2017. Stroudsburg: The Association for Computational Linguistics, p. 33.

Erjavec, T./Krek, S./Fišer, D./Ledinek, N. (2011): Project JOS: linguistic annotation of Slovene. Institut Jožef Stefan, Odsek za tehnologije znanja. http://nl.ijs.si/jos/ (last access: 23-03-2018).

Erjavec, T./Krek, S./Arhar, Š./Fišer, D./Ledinek, N./Saksida, A./Sivec, B./Trebar, B. (2010): Oblikoskladenjske specifikacije JOS V1.1 2010-03-07. http://nl.ijs.si/jos/msd/html-sl/index.html (last access: 23-03-2018).

Fillmore, CH. J./Johnson, Ch. R./Petruck, M. R. L. (2003): Background to Framenet. In: International Journal of Lexicography 16 (3), pp. 235–250.

Gantar, P. (2020): Dictionary of modern Slovene: from Slovene lexical database to digital dictionary database. In: Rasprave Instituta za hrvatski jezik i jezikoslovlje 46 (2), pp. 589–602.

Gantar, P./Štrkalj Despot, K./Krek, S./Ljubešić, N. (2018): Towards semantic role labeling in Slovene and Croatian. In: Fišer, D./Pančur, A. (ed.): Proceedings of the Conference on Language Technologies & Digital Humanities, Ljubljana, September 20–21, 2018. Ljubljana, p. 93.

Gantar, P./Krek, S./Kosem, I./Šorli, M./Kocjančič, P./Grabnar, K./Yerošina, O./Zaranšek, P./Drstvenšek, N. (2013): Slovene lexical database 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1030 (last access: 23-03-2022).

Hanks, P./Pustejovsky, J. (2005): A pattern dictionary for natural language processing. In: Revue Française de linguistique appliquée 10 (2), pp. 63–82.

Hanks, P. (2004): Corpus pattern analysis. In: Williams, G./Vessier, S. (eds.): EURALEX 2004. Proceedings of the Eleventh EURALEX International Congress, Lorient, July 6–10, 2004. Lorient, p. 87.

Keber, J. (2011): Dictionary of Slovenian phrasemes. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1129 (last access: 23-03-2022).

Kosem, I./Pori, E. (2021): Slovenske ontologije semantičnih tipov: samostalniki. In: Kosem, I. (ed.): Kolokacije v slovenščini. 1. izd. Ljubljana, Znanstvena založba Filozofske fakultete, pp. 159–202.

Kosem, I./Krek, S./Gantar, P. (2020): Defining collocation for Slovenian lexical resources. In: Kosem, I./Ganatar, P. (eds.): Collocations in lexicography: existing solutions and future challenges, Slovenščina 2.0 8 (2). Ljubljana, pp. 1–27.

Krek, S./Arhar Holdt, Š./Erjavec, T./Čibej, J./Repar, A./Gantar, P./Ljubešić, N./Kosem, I./Dobrovoljc, K. (2020a): Gigafida 2.0: the reference corpus of written standard Slovene. In: Calzolari, N. (ed.): LREC 202: Twelfth International Conference on Language Resources and Evaluation, Marseille, May 11–16, 2020. European Language Resources Association (ELRA), p. 3340.

Krek, S./Erjavec, T./Dobrovoljc, K./Gantar, P./Arhar Holdt, Š./Čibej, J./Brank, J. (2020): The ssj500k training corpus for Slovene language processing. In: Fišer, D./Erjavec, T. (eds.): Jezikovne tehnologije in digitalna humanistika. Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, September 24–25 2020. Ljubljana: Institute of Contemporary History, p. 24.

Krek, S./Kilgarriff, A. (2006): Slovene word sketches. In: Erjavec, T./Žganec Gros, J. (eds.): Language technologies. Proceedings of the 9th International Multiconference Information Society IS 2006, Ljubljana, 9–10 October 2006. Ljubljana: Institut “Jožef Stefan”, p. 62.

Ramisch, C./Savary, A./Guillaume, B./Waszczuk, J./Candito, M./Vaidya, A./ Barbu Mititelu, V./ Bhatia, A./Iñurrieta, U./ Giouli, V./Güngör, T./Jiang, M./Lichte, T./Liebeskind, Ch./Monti, J./Ramisch, R./Stymne, S./Walsh, A./Xu, H. (2020): Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In: Markantonatou, S./McCrae, J./Mitrović, J./ Tiberius, C./Ramisch, C./Vaidya, A./Osenova, P./Savary, A. (eds.), Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, Barcelona (Online), December 2020. Barcelona, p. 107.

Savary, A./Cordeiro, S. R./Ramisch, C. (2019): Without lexicons, multiword expression identification will never fly: a position statement. In: Savary, A./Parra Escartín, C./Bond, F./Mitrović, J./Barbu Mititelu, V. (eds.): Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Florence, August 2nd, 2019. Florence, p. 79.

Tavast, A./Langemets, M./Kallas, J./Koppel, K. (2018): Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In: Krek, S./Čibej, J./Gorjanc, V./Kosem, I. (eds.): Lexicography in Global Contexts. Proceedings of the 18th EURALEX International Congress, Ljubljana, 17–21 July 2018. Ljubljana, p. 749.

Tiberius, C./Krek, S./Depuydt, K./Gantar, P./Kallas, J./Kosem, I./Rundell, M. (2021): Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources. In: Kosem, I./Cukr, M./Jakubíček, M./Kallas, J./Krek, S./Tiberius, C. (eds.): eLex 2021. Proceedings of the eLex 2021 Conference, virtual, Brno, 5–7 July 2021. Brno, p. 56.

Contact information

Polona Gantar
University of Ljubljana
apolonija.gantar@guest.arnes.si

Simon Krek
Jožef Stefan Institute
simon.krek@guest.arnes.si

Acknowledgements

This paper was written within the framework of the research project New Grammar of Modern Standard Slovene: resources and methods (J6-8256) and in the framework of the research programmes Slovene Language – Basic, Contrastive and Applied Research (P6-0215) and Language Resources and Technologies for the Slovene Language (P6-0411), funded by the Slovenian Research Agency.