1.01 UDC: 003.295:821.163.6'367.625

Polona Gantar,[*] Špela Arhar Holdt,[**] Jaka Čibej,[***] Taja Kuzman[****]

Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene

IZVLEČEK

STRUKTURNA IN POMENSKA KLASIFIKACIJA GLAGOLSKIH VEČBESEDNIH ENOT V SLOVENŠČINI

Prispevek je nadgrajena različica konferenčnega prispevka, v katerem predstavljamo kategorije glagolskih večbesednih enot (GVBE), kot so bile oblikovane v okviru mednarodne COST akcije PARSEME Shared Task 1.1. S kategorijami, ki so nadjezikovne in obenem prilagojene posameznim vključenim jezikom, smo označili 13.511 povedi učnega korpusa ssj500k 2.0. Rezultat označevanja je 3.364 identificiranih večbesednih glagolskih enot, ki so klasificirane kot: inherentno povratni glagoli, zveze z glagoli v pomensko oslabljeni rabi, predložnomorfemski glagoli in glagolski idiomi. V prispevku rezultate označevanja predstavimo kvantitativno in kvalitativno, pri čemer sopostavimo predlagani sistem klasifikacije ob obstoječe prakse na področju slovenistične obravnave GVBE in ocenimo uporabnost sistema za nadaljnje delo.

Ključne besede: glagolske zveze, korpusni pristop, večbesedne enote, PARSEME, slovenščina

[*] Department of Translation, Faculty of Arts, University of Ljubljana, Aškerčeva 12, SI-1000 Ljubljana, apolonija.gantar@guest.arnes.si

[**] CJVT, Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, spela.arhar@cjvt.si

[***] Artificial Intelligence Laboratory, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, jaka.cibej@ijs.si

[****] kuzman.taja@gmail.com

ABSTRACT

This paper is an extended version of a conference paper presenting the categorization of verbal multi-word expressions (VMWEs) according to the PARSEME COST Action Shared Task 1.1 Guidelines. The categorization is universal but takes into account the characteristics of the individual languages included in it. The Shared Task was used to annotate over 13,000 sentences of the Slovene ssj500k 2.0 training corpus, which resulted in nearly 3,400 identified VMWEs categorized as inherently reflexive verbs, light verb constructions, inherently adpositional verbs, and verbal idioms. The paper presents both the quantitative and qualitative results of the analysis, compares the suggested categorization system to existing work on VMWEs in Slovene linguistics, and evaluates the use of the proposed system for future work.

Keywords: verb phrases, corpus approach, multi-word expressions, PARSEME, Slovene

Introduction

In the digital medium, the bulk of interactions between users – as well as between users and computers or applications – occur with the use of language, which is why the existence and open accessibility of digital language infrastructure is of vital importance to the development and vitality of a language. Slovene is no exception in this regard; it requires an infrastructure that serves as a source of information for the language community as well as a resource to be used in applied/theoretical linguistic research and the development of new language technology tools, methods, and services. Examples of such infrastructure include digital language resources that allow for continued updates and contributions from the community, language databases with structured and machine-readable data, and training corpora in which authentic texts are annotated with different linguistic categories. In this regard, digital lexicography, whose aim is to prepare the dictionary part of this language infrastructure, plays an important role in the field of digital humanities.

In the field of digital lexicography, multi-word expressions (MWEs) are considered important for constructing machine-readable language resources that in turn enable the compilation of electronic MWE lexicons and the development of language technology tools for a specific language. In order to achieve these goals, it is crucial to know the linguistic features of different types of MWEs and develop methods and standards for their identification in authentic language use.

However, this is not a trivial task. Definitions and categorisations of MWEs differ according to their methodological and theoretical basis and research goals.[1] A lexicographic perspective focuses on the semantic characteristics of MWEs and defines them

[1] For an overview of MWE classifications according to different methodological approaches, see Gantar et al. (2018).

as "different types of phrases that demonstrate a certain degree of idiomatic meaning" (Atkins and Rundell 2008, 166) or as phrases whose "exact meaning is not directly obtained from its component parts" (Sag et al. 2002). On the other hand, the definition of MWEs from the perspective of machine processing emphasizes their statistical significance: "a group of tokens in a sentence that cohere more strongly than ordinary syntactic combinations: that is, they are idiosyncratic in form, function, or frequency" (Schneider et al. 2014) and their inability to be split into independent lexemes and at the same time maintain their semantic and syntactic functions, as well as their lexical, syntactic, semantic, pragmatic and statistical idiomaticity (Baldwin and Kim 2010, 3). Although no universally accepted definition of MWEs exists, researchers in linguistics and NLP both agree that the key feature separating MWEs from free phrases is the special relation between the elements that form the MWE. This relation is usually defined using such concepts as collocability (or statistical idiomaticity), idiomaticity (or semantic non-compositionality), syntactic (in)flexibility, which also includes the possibility of an internal modification of the phrase and the flexible order of its lexicalized elements, and lexical variance.

An attempt to provide the much needed guidelines and a pilot study on the annotation of MWEs in language corpora was made as part of the PARSEME COST Action Shared Task 1.1.[2] The task focused on the automatic identification of verbal multi-word expressions (VMWEs) in running text. As part of the task, universal guidelines for VMWE classification were compiled containing examples for all languages involved. Based on the guidelines, a multi-lingual corpus was manually annotated with VMWEs and made available under a Creative Commons licence.

While the categories of MWEs were designed as language-independent, the specific characteristics of all the included languages had to be taken into account to reach a solution that was universally applicable. In this paper, we focus on the Slovene results, which will be useful when compiling a digital lexicon of Slovene MWEs, as well as other language resources such as the Dictionary of Modern Slovene (Gorjanc et al. 2017) and a corpus-based grammar of Slovene. The topic was presented in Gantar et al. (2018) with a focus on MWEs and their theoretical framework in Slovene studies. This paper focuses on MWEs from the perspective of a unified concept that was applied to 20 different languages within the PARSEME Shared Task 1.1. A comparison of the results can be found in Ramisch et al. (2018).

Identifying and Categorizing Verbal Multi-Word Expressions

The verb plays a crucial role in the sentence in terms of co-text organization, which is why the PARSEME Shared Task focused on verbal multi-word expressions

[2] *Home – PARSEME*, http://www.parseme.eu.

(VMWEs). For further analysis, it is crucial to determine the differences between the definitions and categorizations of VMWEs as established in Slovene studies on one hand, and in the international PARSEME COST Action on the other. The aim of our task is to adapt the international annotation scheme in order to include Slovene. Our research question focuses on the applicability of the PARSEME system to authentic Slovene texts. Can the adapted PARSEME categories be applied in practice? Are they attributable, robust, one-dimensional, and represented in actual language use? What information do they entail (e.g. in terms of syntax), how can they contribute to the development of new automatic extraction methods, and finally, which problems arise when applying the system to text? In the following sections, we present the annotation method. This is followed by a quantitative and qualitative analysis. The latter is focusing on individual categories, their characteristics, and the potential problems of the approach.

Verbal Multiword expressions – Slovenian case

In Slovene studies, MWEs are divided into a) phraseological units (PUs), in which at least one component carries meaning that differs from one of its denotative "dictionary" senses, and expresses figurativeness, and b) all other multiword expressions (i.e. fixed expressions), which are characterized by a certain degree of fixedness and denote a meaning that can be predicted from the meanings of their elements. PUs are further divided by syntactic structure: the clausal type (which also includes proverbs) and the phrasal type (all non-verbal PUs). In Slovene linguistic theory, verbal MWEs are determined by their morphosyntactic features (Toporišič 1973/74; Kržišnik 1994): a MWE is classified as a VMWE if it includes a verbal element and if it functions as a predicate. However, it remains unclear how to classify examples in which the verbal MWE does not function as a predicate, e.g. *hočeš nočeš* 'like it or not', which includes two verbal elements, but functions as an adverbial.

The problem of categorizing MWEs according to their morphological structure and syntactic function was resolved in PARSEME shared task through the definition that the main criterion for VMWEs is that their syntactic head in the prototypical form is a verb, regardless of the fact whether it can or cannot fulfil other syntactic roles. In addition, Slovene categorizations have so far never treated verbs with the *se/si* morpheme as a separate MWE category. Phrasal verbs that consist of a verb and a preposition and carry an independent meaning were categorized as MWEs only conditionally (Kržišnik 1994, 58).

Verbal Multiword expressions within the Parseme Shared task 1.1

For the categorization of VMWEs within the Parseme Shared task 1.1, exhaustive guidelines[3] were prepared in which the VMWE categories are defined by semantic and syntactic features and are described with decision trees. The identification and categorization process consisted of three steps. In the first step, we identified potential VMWEs consisting of a verb as the syntactic head of the phrase and at least one other word. In the second step, we identified the lexicalized elements within the phrase. In the third step, we used detailed linguistic tests consisting of generic and specific language criteria to determine the correct category of the identified VMWE.

Based on the guidelines, VMWEs are further divided into two classes based on whether the category can be applied to the majority of languages included in the task, or whether they are typical of individual (groups of) languages. The universal categories include verbal idioms (VID) and light verb constructions (LVC), which are further divided into full (LVC.full) and causal (LVC.cause). The quasi-universal categories, which are used within individual groups of languages, include inherently reflexive verbs (IRV), which are typical of most Slavic languages, and verb-particle constructions (VPC), typical of Germanic languages. In the second version of the guidelines, an additional quasi-universal category was added: inherently adpositional verbs (IAV), which require an open syntactic slot and are typical of Slovene and several other Slavic languages.

For Slovene, examples of VMWEs can be found for all the categories with the exception of VPC. For certain categories, however, there are specific characteristics based on syntactic or morphological features of Slovene or on grammatical categories that are generally accepted in Slovene but differ to some extent from other languages. The specific Slovene features will be described along with individual VMWE types.

The Corpus and Annotation Tool

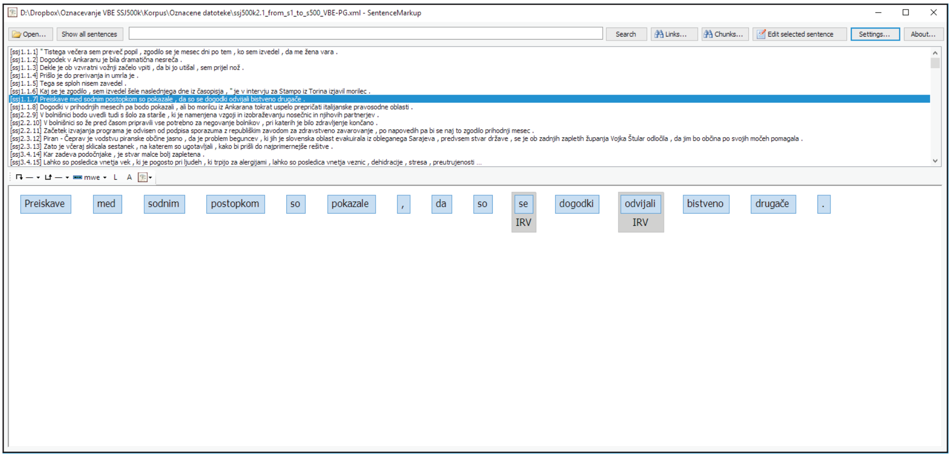
VMWEs were annotated in the Slovene ssj500k 2.0 training corpus (Krek et al. 2017), which consists of approximately 500,000 tokens and just under 28,000 sentences sampled from the FidaPLUS corpus of Slovene (Arhar Holdt and Gorjanc 2007). The entire corpus is morphosyntactically tagged (Grčar et al. 2012). Certain portions also contain named-entity annotations and syntactic dependencies (Dobrovoljc et al. 2012). In the first annotation phase, 11,411 sentences were annotated by two annotators with VMWEs as defined by the first version of the PARSEME Guidelines (Candito et al. 2016). Disagreements in annotation were discussed and adjusted accordingly. In the second phase, the categories were automatically modified based on the second version of the PARSEME Guidelines and manually checked. The

[3] *PARSEME Shared Task 1.1 - Annotation guidelines*, http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/.

second phase continued with the annotation of an additional 2,100 sentences annotated in packages by individual annotators. Problematic examples were discussed and correctly annotated.

The tool used for annotation in the first phase was SentenceMarkup System (Figure 1), a custom tool primarily developed for syntactic dependency annotation of Slovene texts (Dobrovoljc et al. 2012). The tool was adjusted for the annotation of VMWEs by adding an additional independent and interconnectable annotation layer (cf. Gantar et al. 2017).

Figure 1: Annotations in the SentenceMarkup System

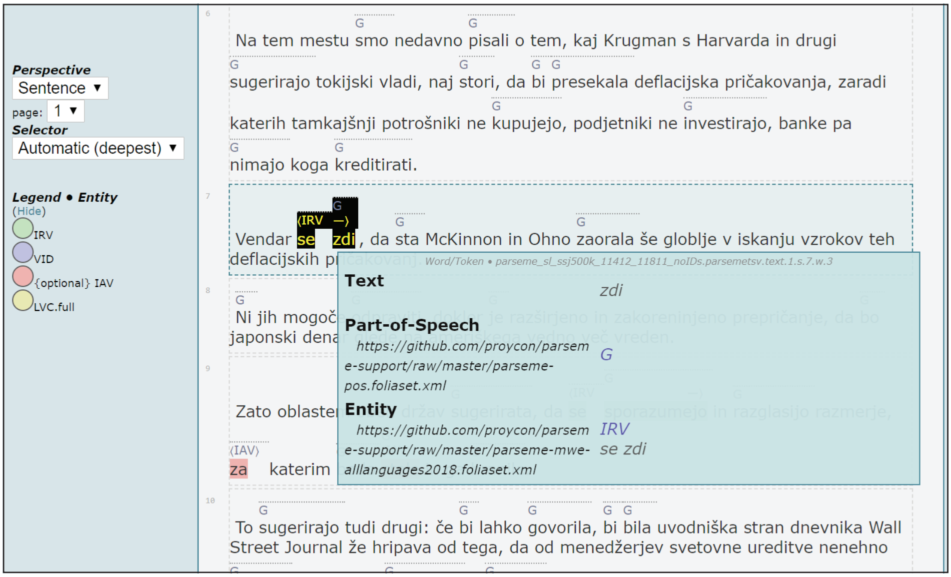

D:\Dropbox\Oznacevanje VBE SSJ500k\Korpus\Oznacene datoteke\ssj500k2.1_from_s1_to_s500_VBE-PG.xml - SentenceMarkup
Open...
Show all sentences
Search
Links...
Chunks...
Edit selected sentence
Settings...
About...
mwe
L
A

[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 vzvratni vožnji začelo vpiti , da bi jo utišal , sem prijel nož .
[ssj1.1.4] Prišlo je do prerivanja in umrla je .
[ssj1.1.5] Tega 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 morilcu 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 pomagala .
[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 ...

Preiskave med sodnim postopkom so pokazale , da so se dogodki odvijali bistveno drugače .

In the second phase, the annotation took place in the FLAT annotation platform (FoLiA Linguistic Annotation Tool), which was adapted for the purposes of the PARSEME Shared Task and tested on 13 collaborating languages (Figure 2). The FLAT platform allows text strings to be annotated with a set of categories. Files can be assigned to different annotators. The supported formats are XML and TSV, while annotated files are exported in XML. All annotations are saved automatically. The interface also features text search using CQL.

Figure 2: Annotations in FLAT



Perspective: Sentence ▾
page: 1 ▾
Selector: Automatic (deepest) ▾

6 Na tem mestu smo nedavno pisali o tem, kaj Krugman s Harvarda in drugi sugerirajo tokijski vladi, naj stori, da bi presekala deflacijska pričakovanja, zaradi katerih tamkajšnji potrošniki ne kupujejo, podjetniki ne investirajo, banke pa nimajo koga kreditirati.

7 Vendar se zdi, da sta McKinnon in Ohno zaorala še globlje v iskanju vzrokov teh deflacijskih pričakovanj.

Word/Token • parseme_sl_ssj500k_11412_11811_noIDs.parsemetsv.text.1.s.7.w.3

Text	zdi
Part-of-Speech https://github.com/proycon/parseme-support/raw/master/parseme-pos.foliaset.xml	G
Entity https://github.com/proycon/parseme-support/raw/master/parseme-mwe-alllanguages2018.foliaset.xml	IRV se zdi

8 Ni jih mogoče odpraviti, dokler je razširjeno in zakoreninjeno prepričanje, da bo japonski denar glede na ameriškega vedno več vreden.

9 Zato oblasteh … držav sugerirata, da se sporazumejo in razglasijo razmerje, za katerim …

10 To sugerirajo tudi drugi: če bi lahko govorila, bi bila uvodniška stran dnevnika Wall Street Journal že hripava od tega, da od menedžerjev svetovne ureditve nenehno

Quantitative Analysis

The annotated VMWEs were imported into the ssj500k 2.1 training corpus (Krek et al. 2018). Among the 13,511 sentences annotated in the first two annotation phases, 2,290 of them (approximately 22%) contain at least one VMWE. On average, each of these sentences features 1.15 VMWEs. Taking into account all the annotated sentences, each sentence contains approximately 0.25 VMWEs; in other words, on average, one VMWE is present in every fourth sentence.

Table 1 shows the distribution of the annotated VMWEs by category. The final number of VMWEs in the training corpus is 3,364. The number of different VMWEs (i.e. without any repetitions of the same unit) was just under 1,100. When looking at absolute frequencies, the most frequent category is IRV (48%) and the least frequent category is LVC.cause (2%). The categories with the highest number of different VMWEs are VID and IAV, while LVC.full and LVC.cause are the least diverse categories. We describe each category in more detail in section 5.

Table 1: Distribution of annotated VMWEs by category

Category	**Example**	**Translation**	**All VMWEs**	**Percent**	**Different VMWEs**
Inherently Reflexive Verbs (IRV)	*bati se*	to be afraid	1,627	48%	345
Inherently Adpositional Verbs (IAV)	*priti do*	to come about	710	21%	154
Verbal Idioms (VID)	*spati kot ubit*	(lit.) to sleep like a dead person	724	22%	457
Light Verb Constructions (LVC): LVC.cause	*spraviti koga v smeh*	to make someone laugh	64	2%	27
Light Verb Constructions (LVC): LVC.full	*biti v pomoč*	to be of help	239	7%	103
Total	-	-	3,364	100%	1,086

Table 2 shows the most common VMWE structures by parts of speech (V – verb, N – noun, A – adjective, R – adverb, Pre – preposition, Pro – pronoun). The structures occurring in the corpus with a frequency below 10 have been categorized as Other. The most frequent structures are V + Pro, V + Pre, V + N and V + Pre + N. Collectively, they account for approximately 85% of all annotated VMWEs.

Table 2: Distribution of annotated VMWEs by part-of-speech structure

Structure	**Example**	**Translation**	**Frequency**	**Percent**
V + Pro	*bati se*	to be afraid	1,663	49%
V + Pre	*priti do*	to come about	535	16%
V + N	*imeti odnos*	to have a relationship	372	11%
V + Pre + N	*biti pod vtisom*	to be under the impression	303	9%
V + Pro + A	*biti si edini*	to be unanimous	146	4%
V + R	*biti res*	to be true	136	4%
V + Pro + Pre + N	*ujeti se v past*	to get caught in a trap	24	1%
V + A	*biti jasno*	to be clear	20	1%
V + A + N	*imeti glavno besedo*	to have the last word	19	1%
N + V + Pre + N	*biti na robu propada*	to be on the verge of collapse	12	<1%
V + Pro + N	*vzeti si čas*	to take one's time	11	<1%
Other	-	-	123	4%
Total	-	-	3,364	100%

Qualitative Analysis

The qualitative analysis deals with the semantic and structural features of VMWEs. Based on the PARSEME Guidelines, several characteristic features of Slovene were identified on the level of structural and semantic tests used to determine the category of VMWEs. In the analysis, we focused on patterns within structures for each subcategory, the syntactic environment of the expression as a unit, and the lexical units filling the corresponding participant slots. Based on corpus examples, we also tried to identify the indicators of semantic integrality that could be useful when automatically identifying VMWEs in text.

Inherently Reflexive Verbs (IRV)

The PARSEME Shared Task 1.1 guidelines treat verbs occurring with the independent morpheme *se/si* as a separate category of VMWEs called *inherently reflexive verbs*. It is a language-specific category that includes phrases in which the verb without the morpheme *se/si* does not exist (*zdeti se* 'to seem', **zdeti*) or in which the presence of *se/si* changes the meaning of the verb (*pobrati se* 'to recover' vs. *pobrati* 'to pick up').

Inherently reflexive verbs cover the largest percentage of VMWEs in the training corpus (see Table 1). Among the correctly categorized examples (1,621 in total)[4] we identified 339 different IRVs, with the following most frequently occurring verbs: *zdeti se* 'to seem', *odločiti se* 'to decide', *zgoditi se* 'to come to pass' and *pojaviti se* 'to appear'.

To test whether the expression is semantically integral and to differentiate it from other types of verb phrases with *se/si* that are not defined as VMWEs, we examined the behavior of the verb in terms of its opening up syntactic positions as a phrase. Inherently reflexive verbs keep *se/si* as an obligatory verb morpheme in all forms of their inflectional paradigm and can be transitive (*bati se koga/česa* 'to be afraid of smn/sth') or intransitive (*znajti se* 'to find oneself somewhere', *zvečeriti se* 'to fall [evening]').

Inherently reflexive verbs as VMWEs must be differentiated from verbs where the reflexive pronoun *se/si* is not an obligatory morpheme but serves another function, more specifically: (a) it denotes mutualness (*poljubljati se* 'to kiss [each other]', *srečati se* 'to encounter [each other]'), (b) it denotes that the target of the action is the subject (*umivati se* 'to wash [oneself]', *praskati se* 'to scratch [oneself]'), or that the action is to the benefit of the subject (*kuhati si* 'to cook [oneself sth]', (c) it is used for passivizing the sentence by removing the agent (*kdo ponavlja kaj* 'someone repeats something' – *kaj se ponavlja* 'something is repeated'), and (d) it denotes a generic action (*govori se* 'it is said'; *se razume* 'it is understood').

With verbs that can also occur without *se/si*, only the phrases where the morpheme changes the verb's meaning are categorized as IRVs. There are cases in which

[4] Among the 1,627 annotated examples, four were mis-categorized. In two examples, the elements of the expressions were incorrectly annotated. These examples were excluded from further analysis.

the presence (or absence) of *se/si* causes a semantic shift directly tied to a human subject In these cases, the verb denotes a metaphorical meaning *pobrati se* 'to recover': *pobrati* 'to pick sth up'; *gristi se* 'to worry' : *gristi* 'to bite'.

In Slovene linguistics, lexicalized phrases consisting of a verb and the *se/si* morpheme have so far not been treated as a fixed expressions. The main focus has been recognition of the function of the morpheme or the reflexive pronoun in terms of denoting different degrees of agentness or the subject's (un)involvedness, as in the case of the non-singular (*zbrati se* 'to gather') or generic agent (*tiskati se* 'to be printed') (Žele 2012, 44; Toporišič 1982, 244; 2000, 503). The identification of IRVs in text from the perspective of their semantic and syntactic is particularly important for the automatic identification of MWEs. In future lexicons and dictionaries, IRVs should thus be treated either as independent entries or as part of polysemy.

Light Verb Constructions (LVC)

Light verb constructions have been treated from different perspectives by different authors (for an overview, see Soršak 2013). In most definitions, the verbs in LVCs are categorized as something between full verbs and auxiliary verbs, while the expressions that feature them are categorized as a phenomenon between fixed and free expressions. Using existing typologies for Slovene (Toporišič 2000; Žele 1999), Soršak analyzes Slovene LVCs based on the entries in the Dictionary of Standard Slovene (SSKJ). The results highlight that the dictionary often mentions the semantically light use of a verb in places where the use is stylistically marked, most frequently as expressive (Soršak 2013, 514; e.g. *groza ga sprehaja,* lit. 'terror is walking him'). The results described in this paper show the opposite – in the annotated corpus, LVCs are typical, stylistically neutral, and frequently occurring.

As per the PARSEME Guidelines, a LVC must fulfil the following conditions: it consists of a verb and a noun or a noun phrase that can also take the form of a prepositional phrase (*imeti mnenje* 'to have an opinion', *biti v dvomih* 'to be in doubt'), and must open up its own valency slots (*kdo ima predavanje za koga* 'someone holds a lecture for someone'). Semantically, the expression must denote an action (*imeti predavanje* 'to hold a lecture') or a state (*biti v dvomih* 'to be in doubt'). According to the verb, the category has two subtypes: (a) if the verb contributes to the meaning on a predominantly categorical level, the expression is categorized as LVC.full (*biti v pomoč* 'to be of help'); (b) if the subject can be interpreted as the cause or source of the denoted action, the expression is categorized as LVC.cause (*spraviti v smeh* 'to make smn laugh'). The LVC tests also take into account the abstractness of the noun (*imeti avto* 'to have a car' is not a multiword expression, while idiomatic expressions like *imeti mačka* 'lit. to have a cat – to have a hangover' are categorized as VIDs) and, with LVC.full, the possibility of rephrasing by omitting the verb (*Janez ima predavanje* 'Janez holds a lecture' – *Janezovo predavanje* 'Janez's lecture').

Despite the somewhat elusive concept of LVCs, the annotation process has confirmed that the PARSEME guidelines are specific enough to be successfully applied to real text. Of the 303 examples annotated as LVCs (1 example was categorized incorrectly), 78.8% were LVC.full and 21.2% LVC.cause. 87.1% of them were combinations of a verb and a noun, while 12.9% were combinations of a verb and a prepositional phrase. The annotated LVCs contained a total of 19 different verbs,[5] predominantly the verb *imeti* 'to have' (65.6%), but also *biti* 'to be' (13.6%) and *dati/dajati* 'to give' (a total of 9.6%).[6] Other verbs (*narediti* 'to do', *postaviti/postavljati* 'to put', *ostati* 'to remain', *voditi* 'to lead', *namenjati* 'to pay [attention]', *delati* 'to do/make', *storiti* 'to do', *vzbujati/zbujati* 'to incite', *dobiti* 'to get', *zastaviti* 'to pose', *spraviti* 'to make', *doseči* 'to achieve' and *nositi* 'to wear') occur less frequently, often in a single expression (*ostati v spominu* 'to remain in one's memory', *namenjati pozornost* 'to pay attention to sth').

Combinations of a verb and a prepositional phrase are somewhat more typical of the LVC.cause category. In the annotated data, LVC.cause occurs exclusively with the prepositions *v* 'in' (33 instances) and *na* 'on' (6 instances). In the majority of cases, the combination is *biti v* (*biti v pomoč* 'to be of help', *biti v podporo* 'to provide support', *biti v navadi* 'to be a habit').

In the annotated expressions, a relatively limited set of nouns can be found: a total of 97. The most frequent nouns are *težava* 'problem' (21) and *pravica* 'right' (20), followed by *možnost* 'possibility', *mnenje* 'opinion', *učinek* 'effect', *vloga* 'role', *vpliv* 'influence', *vtis* 'impression', *pomoč* 'help', *občutek* 'feeling', *prednost* 'advantage', *sreča* 'luck', *korist* 'benefit', *vprašanje* 'question', *volja* 'will', *posledica* 'consequence'. As expected, some of these nouns occur exclusively in LVC.full (*pravica*, *možnost*, *mnenje*, *vloga*), while others occur in LVC.cause (*učinek*, *vpliv*, *vtis*, *pomoč*). In other cases, the category depends on the meaning of the verb (*dati prednost* 'to give an advantage' ® LVC.cause and *imeti prednost* 'to have an advantage' ® LVC.full.

In accordance with the conclusions made by Soršak (2013, 519), the results show that the featured verbs can also be used with full meaning, while the semantic lightness in LVCs is complemented by the nominal part (*imeti* 'to have' meaning 'to possess' compared to *imeti posledice* 'to have consequences' meaning 'to cause/lead to consequences'). Semantically, the noun groups occurring in LVC.cause describe the result of an action, be it a type of result (*učinek* 'effect', *vpliv* 'influence', *vtis* 'impression'), a positive (*korist* 'benefit', *užitek* 'pleasure') or negative consequence (*muka* 'torment', *preglavica* 'trouble'). The semantically light verb binds the result to the subject (*nekdo/nekaj daje vtis* 'smn/sth makes an impression', i.e. the agent is the cause of the action). In certain cases, LVCs can be converted into semantically full verbs with a similar morphological base (*dosegati učinek* 'to achieve an effect' – *učinkovati* 'to affect'; *imeti*

[5] This is the full set of the LVCs in the data, confirming that the set of verbs occurring in these expressions is limited. In the dictionary, Soršak (2013, 513) finds mentions of semantic lightness in 420 verb entries. However, as mentioned, the labels often signify stylistically marked and atypical language use.

[6] In Slovene lingustics, verb phrases with *imeti* 'to have' and *biti* 'to be' have been most frequently treated as the equivalent of LVCs, but analyzed from different perspectives (see e.g. Vidovič Muha 1998).

vpliv 'to have an influence' – *vplivati* 'to influence'), but not always (*imeti posledice* 'to have consequences' – /).

The nouns occurring in LVC.full are semantically more diverse. Dividing them into semantic groups reveals that the common ground of these expressions can be defined as planning or estimating success. Among the encountered LVCs are phrases with nouns dealing with (a) communication (*mnenje* 'opinion', *predlog* 'suggestion', *vprašanje* 'question'); or describing (b) the potential for success (*možnost* 'possibility', *prednost* 'advantage', *priložnost* 'opportunity'); (c) initial steps (*obljuba* 'promise', *napoved* 'prediction', *načrt* 'plan'); (d) potential reasons for failure (*napaka* 'mistake', *pomanjkljivost* 'disadvantage'). Other groups deal with (e) negative states (*težava* 'problem', *strah* 'fear', *dvom* 'doubt'), (f) positive qualities (*moč* 'power', *pogum* 'courage', *potrpljenje* 'patience'), (g) achieved results (*izobrazba* 'education', *status* 'status', *posel* 'business'), and (h) attitude towards as of yet unrealized goals (*želja* 'wish', *ambicija* 'ambition', *vizija* 'vision'). Again, some examples can be converted into a semantically full verb (*imeti mnenje* 'to have an opinion' – *meniti*), while others cannot (*imeti ambicije* 'to have ambitions' – /).

Inherently Adpositional Verbs (IAV)

Inherently adpositional verbs, also called verbs with a lexicalized prepositional morpheme (Žele 2002), were included in the PARSEME Guidelines during the second annotation phase as an optional test category.[7] The guidelines define IAVs as verbs that only occur with a prepositional morpheme (*simpatizirati z* 'to sympathize with') or verbs that change meaning when occurring with a prepositional morpheme (*biti za* 'be for, to support' vs. *biti* 'to be'). The participants required by the verb phrase as a whole are not a part of the VMWE, as opposed to e.g. *stati na* + *trdnih tleh* 'to stand on + solid groud', which is categorized as a VID.

Prepositions have been treated as free verb morphemes as early as in Metelko's Grammar of Slovene (1825, 247–56) and were analyzed in further detail by Breznik (1916, 250; 1934, 225). Verbs with a lexicalized prepositional morpheme were also analyzed by Žele (2002) and Kržišnik (1994), the former from the perspective of the degree of lexicality of the preposition and the latter from the perspective of phrase fixedness as either a phraseological unit with structural fixedness (*biti ob čem* 'to be next to sth' meaning 'to be positioned next to sth') or phrasemes with lexical fixedness (*biti ob kaj* 'to lose sth').

[7] Based on the feedback from the first annotation campaign and the issues discussed among the contributors, idiomatic combinations of verbs with prepositions or postpositions (IAVs) were separated from verb-particle constructions (VPCs) such as *put off, to blow up, to do in*, in which the particle completely changes the meaning or adds a partly predictable but non-spatial meaning to the verb. Unlike VPCs, which are characteristic of Germanic languages and Hungarian, less so of Romance languages, and absent in Slavic languages, IAVs can exclusively be found in the Balto-Slavic language group.

In the training corpus, IAVs account for approximately 20% of all annotated VMWEs (see Table 1). Among the 710 examples, 154 diverse IAVs were identified. The following examples appear with a frequency of at least 20: *iti za* 'to be about' (always in the third person singular – *gre za*), *priti do* 'to occur', *ukvarjati se z* 'to work on sth', *vplivati na* 'to influence', *skrbeti za* 'to take care of', *temeljiti na* 'to be based on', *naleteti na* 'to encounter', *veljati za* 'to be considered' and *biti proti* 'to be against'. As per the guidelines, the IAV category also includes verb phrases that consist of an inherently reflexive verb (see 5.1) and a lexicalized prepositional morpheme (*nanašati se na* 'to refer to sth').

The most frequent lexicalized prepositional morpheme is *za* 'for', occurring with 34 different verbs (e.g. *gre za* 'to be about'), followed by *na* 'on', occurring with 33 different verbs (e.g. *vplivati na* 'to influence'). Frequent prepositional morphemes are also *z/s* 'with', *do* 'to' and *v* 'in'.

The lexicalized prepositional morpheme is usually positioned after the verb, which is true in 86% of the annotated examples. In the vast majority of cases, the morpheme is positioned directly after the verb or in a narrow window (+3 words). An exception is *gre za*, where an intervening element serves to reference preceding information (*gre [v tem primeru] za* 'it [in this case] is about'). In less frequent examples where the prepositional morpheme is positioned before the verb, the distance between the verb and the morpheme is significantly larger (in 20% of the cases, the distance is 3+ words).

Verbs with a lexicalized prepositional morpheme can also be identified based on common semantic features, e.g. the expression of (a) function or quality: *veljati za* [*favorita*] 'to be considered [a favorite]',[8] *imenovati [direktorja]* 'to name [smn a director]', *označiti za [laž]* 'to call [sth] out as [a lie]'; (b) (dis)agreement: *biti za/proti [globalizacijo]* 'to be for/against [globalization]'; (c) basis: *temeljiti na (dejstvu)* 'to be based on [fact]', *graditi na (zaupanju)* 'to build on [trust]'; (d) beginning or change of action/state: *pasti v [komo]* 'to fall in [a coma]', *prerasti v (ljubezen)* 'to blossom into [love]'; (e) change of quality or form: *pretvoriti v (energijo)* 'to convert into [energy]'; (f) survival: *iti skozi (proces)* 'to go through [a process]'; (g) active participation: *ukvarjati se z* 'to work on sth', *skrbeti za* 'to take care of sth'.

IAVs are characterized by the fact that the presence of the prepositional morpheme often changes the valency qualities of the verb, e.g. (a) when the original intransitive verb becomes transitive, as in the example *živeti* 'to live' : *živeti od koga/česa* 'to live off of sth'; (b) when there is a change in the case of the prepositional complement, e.g. *obrniti se na koga* 'to turn to someone (fig.) : *obrniti se h komu* 'to turn to someone (lit.)'. There are also many examples of movement verbs that as IAVs change meaning to a non-spatial judgment of state (*priti skozi* 'to go through' in the sense of 'to survive'). With verbs featuring a wide semantic range, the prepositional morpheme typically narrows down the meaning (*biti* 'to be' : *biti za* 'to be for, to support sth'). Some verbs within IAVs require an abstract object, e.g. *pasti v [depresijo, vrtinec nizkotnosti]* 'to fall

[8] With IAVs, we also list typical collocates from the Gigafida Corpus of Written Slovene to ease semantic disambiguation.

into [depression, a whirlpool of insidiousness]', *dišati po* [*prevari*] 'to smell of [deceit]', *pokati od* [*veselja*] 'to be bursting of [joy]'.

Identifying inherently adpositional verbs poses a challenge both for human annotators and language technology tools as additional elements can intervene between the lexicalized morpheme and the verb. In addition, numerous verb-preposition combinations can denote a literal meaning while not exhibiting any change in the case of the object complement (*stati za* [*vrati*] 'to stand behind the door' : *stati za* [*dejanji*] 'to stand by one's actions'). They can also be polysemous (*priti do* [*spremembe*] 'to occur [change]' : *priti do* [*denarja*] 'to get [money]'). The analysis offers a starting point for the automatic identification of IAVs and provides possibilities for more detailed research, especially in terms of valency, sentence patterns and the semantic features of participants.

Verbal Idioms (VID)

The PARSEME Guidelines define verbal idioms (VID) as the combination of two lexicalized elements in which the verb is the syntactic head that requires at least one participant in the sentence pattern. The participants can take different syntactic roles, e.g. a direct or prepositional object complement (*plačati ceno* 'to pay a price', *zravnati z zemljo* 'to level with the earth'), a subject (*zgodba se ponavlja* 'lit. the story repeats itself'), an adverbial (*spati kot ubit* 'lit. to sleep like a dead person') or a subordinate clause (*vedeti, koliko je ura* 'lit. to know what time it is' in the sense 'to know what is going on'). VIDs must also keep a meaning that is independent of the meanings of their elements even with certain syntactic conversions. The Guidelines mention that the elements can appear in expected paradigms (declensions), in different tenses, in active or passive voices, with lexical variance, etc.

The definition provided by the PARSEME Guidelines differs from the one found in Slovene linguistics in that it focuses on the verb as the head and the lexicalized elements within the verb's sentence pattern. On the other hand, Slovene linguistics focuses primarily on the ability of the verb phrase as a whole to take the role of the predicate (Toporišič 1973/74; Kržišnik 1994). From this point of view, it is problematic to treat phrases that feature a verb as the fixed part, but as a whole do not always take the role of the predicate. In some cases, they can take the role of an object complement (*[ne spodobi se] voditi za nos* 'lit. [it is not proper] to lead someone by the nose' in the sense 'fooling someone is frowned upon'), a sentence (*srce se trga* [*komu*] '[someone's] heart is breaking'), or an adverbial (*hočeš nočeš* 'like it or not').

In the training corpus, 724 units were categorized as VIDs, which represents 22% of all VMWEs (see Table 1). As can be expected, VIDs occurring more than 10 times feature the verbs *biti* 'to be' and *imeti* 'to have'. Several other VIDs occur more than 5 times (*biti kos* 'to be sth's match', *priti prav* 'to come in handy', *igrati vlogo* 'to play a role', *pustiti pri miru* 'leave sth be', *priskočiti na pomoč* 'to rush to smn's aid', and *imeti opravka* *s/z* ‘to busy oneself with’), along with fixed discourse markers (cf. Dobrovoljc 2017): *se pravi* ‘which is to say’, *kdo ve* ‘who knows’.

As mentioned above, the most frequent structures are combinations of the verb *biti* ‘to be’ and an adverb/adjective/noun. Taking into account their structural fixedness and semantic vagueness of the verb, they should be treated as separate lexicon entries: *biti všeč/res/mar/prida/prav/kos* ‘to be likeable/true/to care/to be of benefit/to be right/to be smn’s match’. This group includes phrases with a semantically wide verb *imeti* ‘to have’: *imeti prav/rad* ‘to be right/to love’, *ne imeti pojma/smisla* ‘to have no clue/meaning’.

Another frequent structure in the training corpus is the combination of a verb and a noun or noun phrase. Among the verbs, the most frequent are *delati* ‘to make’ (*delati družbo/gužvo/izjeme/preglavice/razlike/sceno/škodo* ‘to do/make company/a crowd/an expection/trouble/a difference/a scene/damage’) and *dati* ‘to give’ (*dati polet/pečat* ‘to give momentum/to leave a mark’). The latter structurally coincide with LVCs, but cannot be converted in the same way as LVCs to express possession (*Miha ima predavanje* ‘Miha holds a lecture’ ® *Mihovo predavanje* ‘Miha’s lecture’, but not *Miha dela gužvo* ‘Miha is crowding the place’ ® _*Mihova gužva_ ‘Miha’s crowd’). The largest percentage in the training corpus is covered by VIDs consisting of a verb and a prepositional phrase. Again, the most frequent verb is *biti* ‘to be’ (*biti na dosegu roke* ‘to be in reach’, *biti na razpolago/voljo* ‘to be at one’s disposal’), followed by e.g. *priti* ‘to come’ (*priti na dan* ‘to come to light’, *priti na misel* ‘to come to mind’) and *dati* ‘to give’ (*dati na izbiro* ‘to give a choice’). In terms of fixedness, some combinations of a verb and a nominal/prepositional phrase require an obligatory negation (*ne moči si kaj* ‘can’t help but’, *ni ne duha ne sluha o (kom/čem)* ‘no trace of sth’, *ni para (komu)* ‘someone has no equal’).

The training corpus also features other structures, but with lower frequencies (*solze stopijo v oči (komu)* ‘someone’s eyes are watering’, *časi se spreminjajo* ‘times are changing’). These also include idioms (*bolje preprečiti kot zdraviti* ‘lit. better to prevent than to cure’) and comparisons (*igrati se [s kom/čim] kot mačka z mišjo* ‘lit. to play [with smn/sth] like a cat plays with a mouse’), as well as verb-adverb combinations (*priti skupaj* ‘to come together’, *daleč priti* ‘to come far’) and combinations of a verb and a pronominal morpheme (*zagosti jo (komu)* ‘to create mischief for someone’).

Within their sentence patterns, VIDs open up predictable syntactic slots filled by participants with typical semantic roles. A quick overview of the annotated examples shows that certain verb forms are fixed or more frequent (e.g. third person or negated forms) and that lexical elements in a certain slot are to some extent predictable: *(svet, življenje, vse) postaviti na glavo* ‘to turn [the world/life/everything] upside down’).

Discussion and Conclusion

The conducted annotation task has shown that the annotation set-up (including the tool and the annotation scheme) is suitable. However, content-wise, the task is relatively complex and requires a more advanced linguistic background. The categories provided in the available guidelines are attributable and formalistically distinguishable from each other; categorization problems occur mostly when distinguishing collocations from VMWEs. The quantitative analysis shows that all categories are robust and present in authentic texts.

Based on the annotated VMWEs, we were able to identify certain pattern features on the syntactic and semantic levels. These patterns represent a good starting point for a set of rules for the automatic extraction of VMWEs and for further language description. Methodologically, we made a shift in focus from a functional-syntactic perspective to the description of interconnected features on the morphosyntactic, syntactic, semantic, and lexical levels.

As expected, VMWEs are typically formed by verbs with a wide semantic range, e.g. *biti* 'to be', *dati* 'to give', *imeti* 'to have', which makes them lose their lexical qualities, but keep their morphological features, syntactic function, and position in the sentence pattern. The degree to which the meaning of the verb as an element of the MWE contributes to the meaning of the whole is often difficult to determine, one of the reasons being that numerous verb phrases structurally coincide with several categories, but denote no idiomatic meaning. In text, they are difficult to distinguish from free phrases or collocations (frequent, semantically sensible and structurally adequate word co-occurrences).

On the other hand, the initial structural and semantic analysis has shown that (a) individual types of VMWEs form recognizable structural patterns, e.g. verb + nominal/prepositional phrase; (b) the lexicalization of elements influences the change in the participants' position and their semantic roles (*vreči se po kom* 'to take after smn' – *vreči se v kaj* 'to begin working enthusiastically' – *vreči koga ven* 'to throw smn out'); (c) that the sequence of verb elements in a VMWE is usually not fixed, but (e) there are certain tendencies in word order and (d) the number and representation of intervening elements. Furthermore, (e) certain lexical elements can be predicted based on the frequency and the elements of the co-text; (f) for better automatic identification of VMWEs, their formalized description should include information on all levels of language description.

The list of VMWEs obtained from the annotated corpus represents a set of lexicon units that can be used in machine learning for the automatic identification of VMWEs in text.

While our research did not include a systematic analysis of the sentence patterns, it should be mentioned that the training corpus includes the syntactic (formalized syntactic dependencies) and semantic (semantic role labeling) data that can be used to analyze them. This would allow us to identify more general sentence patterns for a certain VMWE type and use them in automatic extraction.

To correctly identify different MWEs, we will also create a typology of non-verbal MWEs, e.g. nominal (*žlahtna kapljica* 'fine wine'), adjectival (*vreden greha* 'worthy of sin'), or adverbial phrases (*zdaj ali nikoli* 'now or never'), as well as phrases containing particles, conjunctions and pronouns (*ja pa ja* 'as if', *s tem da* 'taking into account that') which were identified as frequent n-grams (Dobrovoljc 2017). Another challenge to tackle is the relation between the canonical and converted forms of MWEs, e.g. začarani krog 'vicious circle' – *biti ujet v začarani krog/v začaranem krogu* 'to be caught in a vicious circle' – *izviti se/rešiti se iz začaranega kroga* 'to escape from a vicious circle' – *vrteti se/znajti se v začaranem krogu* 'to spin/end up in a vicious circle' – *izstopiti iz začaranega kroga* 'to step out of a vicious circle', etc. Furthermore, it is difficult to identify MWEs with an independent, but non-metaphorical meaning, e.g. fixed expressions of the type *tehnološki park* 'technological park' and *ustavno sodišče* 'supreme court', which are closer to terminology and named entities.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency: (a) research core funding No. P6-0215, *Slovene Language – Basic, Contrastive, and Applied Studies*; (b) research core funding No. P6-0411, *Language Resources and Technologies for Slovene*; and (c) project funding No. J6-8256, *New grammar of modern standard Slovene: resources and methods*. The research was conducted within the framework of the IC1207 PARSEME COST Action[9] and the IS1305 ENeL COST Action.[10]

Sources and Literature

- Arhar Holdt, Špela, and Vojko Gorjanc. 2007. "Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa." *Jezik in slovstvo* 52, No. 2 (January): 95–110.
- Atkins, Sue B. T., and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography.* New York: Oxford University Press.
- Baldwin, Timothy, and Su Nam Kim. 2010. "Multiword Expressions" In *Handbook of Natural Language Processing*, edited by Nitin Indurkhya and Fred J. Damerau, Second Edition, 267–92. Boca Raton: CRC Press.
- Breznik, Anton. 1916. *Slovenska slovnica za srednje šole*. Celovec: Družba sv. Mohorja.
- Breznik, Anton. 1934. *Slovenska slovnica za srednje šole*. 4th, enlarged edition. Celje: Družba sv. Mohorja.
- Candito Marie, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Mihaela Ionescu, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Monica-Mihaela Rizea, Agata Savary, Ivelina Stonayova, Sara Stymne, Veronika Vincze. 2016. *PARSEME shared task 1.0 annotation guidelines – version 1.6b* – last updated on November 26, 2016. http://parsemefr.lif.uiv-mrs.fr/parseme-st-guidelines/1.0/.

[9] *Home – PARSEME*, http://www.parseme.eu.

[10] *Action IS1305 – COST*, www.elexicography.eu.

- Dobrovoljc, Kaja. 2017. "Multi-word discourse markers and their corpus-driven identification: the case of MWDM extraction from the reference corpus of spoken Slovene." *International journal of corpus linguistics* 22, No. 4 (December): 551–82.
- Dobrovoljc, Kaja, Simon Krek, and Jan Rupnik. 2012. "Skladenjski razčlenjevalnik za slovenščino." In *Zbornik Osme konference Jezikovne tehnologije*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 42–47. Ljubljana: Jožef Stefan Institute.
- Gantar, Polona, Lut Colman, Carla Parra Escartín and Héctor Martínez Alonso. 2018. "Multiword Expressions: Between Lexicography and NLP." *International Journal of Lexicography*: 1–25.
- Gantar, Polona, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman, and Teja Kavčič. 2018. "Glagolske večbesedne enote v učnem korpusu ssj500k 2.1." In *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Andrej Pančur, 85–92. Ljubljana: Znanstvena založba Filozofske fakultete.
- Gantar, Polona, Simon Krek, and Taja Kuzman. 2017. "Verbal multiword expressions in Slovene." *Europhras 2017, Computational and Corpus-Based Phraseology: proceedings*, edited by Ruslan Mitkov, 247–59. Cham: Springer.
- Godec Soršak, Lara. 2013. "Glagoli z oslabljenim pomenom v Slovarju slovenskega knjižnega jezika." *Slavistična revija* 61, No. 3 (March): 507–22.
- Gorjanc, Vojko, Polona Gantar, Iztok Kosem, and Simon Krek, eds. 2017. *Dictionary of modern Slovene: problems and solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/book/15.
- Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. 2012. "Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik." In *Zbornik Osme konference Jezikovne tehnologije,* edited by Tomaž Erjavec and Jerneja Žganec Gros. Ljubljana: Jožef Stefan Institute.
- Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, and Taja Kuzman. 2017. "Training corpus ssj500k 2.0." *Slovenian language resource repository CLARIN.SI.* http://hdl.handle.net/11356/1165.
- Kržišnik, Erika. 1994. "Slovenski glagolski frazemi (ob primeru glagolov govorjenja)." PhD diss., Faculty of Arts, University of Ljubljana.
- Metelko, Franc Serafin. 1825. *Lehrgebäude der slowenischen Sprache im Königreiche Illyrien und in den benachbarten Provinzen*. Laibach: Leopold Eger.
- Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar et al. 2018. "Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions." In *Proceedings: LAW-MWE-CxG 2018, The 12th Linguistic Annotation Workshop (LAW XII) and the 14th Workshop on Multiword Expressions (MWE 2018)*, edited by Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan, and Miriam R. L. Petruck, 222–40. Santa Fe: Association for Computational Linguistics. http://aclweb.org/anthology/W18-49.
- Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. "Multiword Expressions: a Pain in the Neck for NLP." In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing 2002), edited by Alexander Gelbukh, 1–15. Berlin, Heidelberg, New York: Springer.
- Schneider, Nathan, 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." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014),* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, 455–61. European Languages Resources Association (ELRA).
- *Slovar slovenskega knjižnega jezika*. 2nd edition. Ljubljana: SAZU and Fran Ramovš Institute of the Slovenian Language ZRC SAZU. www.fran.si.
- Toporišič, Jože. 1973/74. "K izrazju in tipologiji slovenske frazeologije." *Jezik in slovstvo* 19, No. 8 (Spring): 273–79.

- Toporišič, Jože. 1982. *Nova slovenska skladnja*. Ljubljana: Državna Založba Slovenije.
- Toporišič, Jože. 2000. *Slovenska slovnica*. Maribor: Založba Obzorja.
- Vidovič-Muha, Ada. 1998. "Pomenski preplet glagolov imeti in biti – njuna jezikovnosistemska stilistika." *Slavistična revija* 46, No. 4: 293–323.
- Žele, Andreja. 1999. "Vezljivost v slovenskem knjižnem jeziku (s poudarkom na glagolu)." PhD diss., Faculty of Arts, University of Ljubljana.
- Žele, Andreja. 2002. "Prostomorfemski glagoli kot slovarska gesla." *Jezikoslovni zapiski* 8, No. 1: 95–108.
- Žele, Andreja. 2012. *Pomensko-skladenjske lastnosti slovenskega glagola*. Linguistica et philologica 27. Ljubljana: Založba ZRC, ZRC SAZU.

Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman

Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene

SUMMARY

In the paper, we present an analysis of Slovene verbal multi-word expressions (VMWEs) based on the categorization made within PARSEME COST Action Shared Task 1.1 for 20 different languages. The purpose of the task was to identify VMWEs in running text based on syntactic and semantic guidelines, as well as to compile a manually annotated multi-language corpus to be made available under a Creative Commons licence. The results of the analysis will be useful in the compilation of a digital lexicon of Slovene multi-word units and will help establish a theoretical framework that takes into account the specific characteristics of Slovene while still fulfilling international criteria.

Unlike the functional-syntactic criteria advocated thus far in Slovene studies (Toporišič 1973/74; Kržišnik 1994), the classification of VMWEs within the PARSEME Shared Task 1.1 focuses on the identification of the syntactic head of the MWE. This allows MWEs to be divided into e.g. verbal, adjectival, and nominal MWEs regardless of the function they have in the sentence as a semantic and syntactic whole. The PARSEME classification consists of both universal and language-specific categories. Universal categories include verbal idioms (VID; *plačati ceno* 'to pay the price') and light verb constructions, which are further divided into full (LVC.full; *imeti mnenje* 'to have an opinion') and causal (LVC.cause; *spraviti v smeh* 'to make smn laugh'). Language-specific categories encompass inherently reflexive verbs (IRV; *zdeti se* 'to seem'), which are typical of most Slavic languages; phrasal verbs (VPC), typical of Germanic languages; and inherently adpositional verbs (IAV), also typical of most Slavic languages, including Slovene. A total of 13,511 sentences in the Slovene training corpus ssj500k 2.0 (Krek et al. 2017) were annotated with 3,364 VMWEs: 1,627 IRV

(48%), 724 VID (22%), 710 IAV (21%), 239 LVC.full (7%), and 64 LVC.cause (2%).

A linguistic analysis of the individual categories highlights numerous semantic and syntactic characteristics of the identified VMWEs that can be taken into account in the compilation of a MWE lexicon and the automatic identification of MWEs in text. Among other things, the results show the importance of the criteria used to distinguish between different types of reflexive verbs based on the role of the reflexive pronoun; they can be viewed either as independent lexical units with their own meaning (e.g. *delati se* 'to pretend') or as verbal phrases denoting e.g. mutual (*poljubljati se* 'to kiss each other'), reflexive (*umivati se* 'to wash oneself'), or passive actions (*ponavljati se* 'to be repeated'). The analysis has also shown that although the order of the components of a VMWE is usually not fixed, certain tendencies exist in terms of word order and the number of intervening elements. A semantic analysis of VMWEs has also revealed the presence of semantic groups formed by VMWEs within an individual category, as well as the properties of light verbs and verbs that typically form idiomatic units.

The study provides a good basis for further analyses of Slovene MWEs. In the training corpus, VMWE annotations can be analyzed in terms of their formalized syntactic dependency trees or the semantic roles played by the participants in the sentence.

Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman

STRUKTURNA IN POMENSKA KLASIFIKACIJA GLAGOLSKIH VEČBESEDNIH ENOT V SLOVENŠČINI

POVZETEK

V prispevku predstavljamo analizo glagolskih večbesednih enot (GVBE) v slovenščini na podlagi kategorizacije, kot je bila izdelana v okviru PARSEME COST Action Shared Task 1.1 za 20 različnih jezikov. Namen naloge je bil identificirati GVBE v tekočem besedilu na podlagi skladenjskih in pomenskih smernic ter izdelava ročno označenega večjezičnega korpusa, ki bo na voljo pod licenco Creative Commons. Rezultati analize bodo uporabljeni pri izdelavi digitalnega leksikona večbesednih enot za slovenščino kot tudi za utemeljitev teoretičnih izhodišč, ki upoštevajo specifike slovenščine in so hkrati usklajena z mednarodnimi merili.

Klasifikacija VMWE znotraj Parseme Shared task 1.1 za razliko od funkcijsko-skladenjskih meril, ki jih predvideva slovenistično jezikoslovje (Toporišič 1973/74; Kržišnik 1994), postavlja v izhodišče prepoznavanje skladenjskega jedra MWE, kar omogoča njihovo delitev na glagolske, pridevniške, samostalniške ipd. GVBE, neodvisno od funkcije, ki jo v stavku opravljajo kot pomenska in skladenjska celota. V izhodišču predvideva Parsemovska klasifikacija univerzalne in jezikovnospecifične

kategorije. Znotraj prvih loči glagolske idiome (VID; *plačati ceno*) in zveze z glagoli v pomensko oslabljeni rabi, ki so členjeni na prave (LVC.full; *imeti mnenje*) in vzročne (LVC.cause; *spraviti v smeh*). Znotraj druge skupine pa inherentno povratne glagole (IRV; *zdeti se*), ki so tipični za večino slovanskih jezikov, frazne glagole (VPC), značilne za germanske jezike, in glagole z leksikaliziranim predložnim morfemom (IAV), ki so tipični za slovenščino in večino slovanskih jezikov. V učnem korpusu ssj500k 2.0 (Krek et al. 2017) smo označili 13,511 stavkov, v katerih smo identificirali skupno 3,364 VMWE v naslednjih deležih: 1,627 IRV (48 %), 724 VID (22 %), 710 IAV (21 %), 239 LVC.full (7 %) in 64 LVC.cause (2 %).

Jezikoslovna analiza posameznih kategorij je pokazala številne semantične in skladenjske značilnosti identificiranih GVBE, ki jih bo mogoče upoštevati pri izdelavi leksikona VBE ter pri njihovi avtomatski identifikaciji v besedilu. Med drugim je izpostavila merila za ločevanje različnih tipov povratnih glagolov na podlagi vloge povratnega zaimka, kar omogoča njihovo obravnavanje bodisi kot samostojnih leksikalnih enot z lastnim pomenom (npr. *delati se*) bodisi kot glagolskih zvez v različnih upovedovalnih vlogah, kot so npr. vzajemnost (*poljubljati se*), povratnost (*umivati se*), pasivizacija (*ponavljati se*) ipd. Analize so tudi pokazale, da zaporedje elementov v GVBE navadno ni ustaljeno, obstajajo pa določene tendence glede besednega reda in števila vrivajočih se elementov. Analiza GVBE s semantičnega vidika je pokazala navzočnost določenih semantičnih skupin, ki jih tvorijo GVBE v posamezni kategoriji, kot tudi lastnosti glagolov v pomensko oslabljeni rabi ter glagolov, ki tipično tvorijo idiomatične enote.

Raziskava postavlja dobre osnove za nadaljnje analize VBE v slovenščini, zlasti ob upoštevanju skladenjskih oznak v obliki formaliziranih skladenjskih drevesnic v učnem korpusu, in semantičnih vlog, pripisanih udeležencem v stavčnem vzorcu.