Verbal Multiword Expressions in Croatian

Krešimir Šojat
Faculty of Humanities and
Social Sciences
University of Zagreb
ksojat@ffzg.hr

Matea Filko
Faculty of Humanities and
Social Sciences
University of Zagreb
msrebaci@ffzg.hr

Daša Farkaš
Faculty of Humanities and
Social Sciences
University of Zagreb
dberovic@ffzg.hr

Abstract

The paper deals with verbal multiword expressions in Croatian. We focus on four types of verbal constructions: light verb constructions, i.e. constructions consisting of a light verb and a noun or prepositional phrase, complex predicate constructions, i.e. constructions consisting of a finite and infinitive verb, prepositional verb constructions, i.e. constructions consisting of a verb and a typical preposition, and, finally, verbal idioms, i.e. constructions with completely idiosyncratic meanings. All the constructions are annotated in the Universal Dependency treebank for Croatian. The identification of verbal multiword expressions is an important task in numerous NLP tasks. It is also important to define and delimitate this concept in linguistic theory.

1. Introduction

The identification and annotation of multiword expressions in Croatian corpora and treebanks have so far gained little attention, although these constructions pose a challenge for various NLP tasks.

Multiword expressions (MWEs) refer to various types of constructions consisting of two or more words that act as a single unit at some level of analysis. Sag et al. (2002) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)" and provide an extensive account of various MWEs in English as well as of criteria for their classification. Generally, MWEs are divided into those that are fixed, i.e. the paradigmatic selection of elements and their syntagmatic order is never altered, and those that can be modified to a certain degree, either in morphosyntactic properties of elements and/or their selection. The meaning of MWEs can vary from more or less compositional to completely idiosyncratic. MWEs usually include noun compounds, multiword named entities, different types of complex verb phrases, idioms and others.

Reporting on annotation schemes in 17 dependency and constituency based treebanks for 15 languages, Rosen et al. (2016:179) point out that there is little agreement on how MWEs should be annotated in treebanks. On top of that, they stress that "there is, in fact, not even agreement on what constitutes a MWE in NLP". Baldwin and Kim (2010) and Rosen et al. (2016) divide MWEs into following groups: 1. nominal MWEs; 2. verbal MWEs; 3. prepositional MWEs; 4. adjectival MWEs; 5. MWEs of other categories; 6. proverbs.

In this paper we focus on verbal MWEs in Croatian. We deal with this type of MWEs because a) there is no previous research done on the identification and annotation of verbal MWEs in Croatian language resources, primarily treebanks and b) there is no resource which would enable an extensive research of MWEs in Croatian and refinement of linguistic criteria for their classification. The paper is structured as follows: In section 2 a brief description of verbal MWE in Croatian is presented and criteria for their classification are given. Section 3 describes the procedure for annotating verbal MWEs and the reasons for selecting the Universal Dependency for Croatian for this purpose. In sections 4 and 5 the results obtained by the MWE annotation of Universal Dependency treebank are presented and discussed. The paper ends with concluding remarks and an outline of future work.

2. Verbal MWEs in Croatian

Both Baldwin and Kim (2010) and Rosen et al. (2014, 2016) divide verbal MWEs into subgroups of phrasal verbs, light verb constructions, VP idioms and other verbal MWEs. All these subgroups require a careful examination for Croatian.

The category of phrasal verbs is generally neither recognized nor discussed in Croatian grammars and reference books. However, Katunar et al. (2012) point out that a particular preposition can significantly change the meaning of a verb and argue that such expressions should therefore be treated as a single unit. The meaning of a verb that co-occurs with an object PP can be significantly different from the meaning of the same verb that co-occurs with an adverbial PP. For instance: 1. *zagrijati se pod pokrivačem* 'to warm up under the blanket' vs. 2. *zagrijati se za kuhanje* 'to become interested in cooking', where the meaning of the verb *zagrijati se* is completely different when used with different PPs.

Light verb constructions (e.g. *donijeti odluku* 'to make a decision; to reach a decision') are made up of a verbal and a nominal component. The nominal component consists of a NP or a PP. Noun in NPs are generally derived from verbal stems and they are usually in accusative case. Light verbs have entirely or partially lost their lexical meaning and the meaning of the whole construction is actually expressed by NPs or PPs. Light verb constructions (LVCs) in Croatian are syntactically flexible since light verbs can be inflected, passivized and marked as perfectives or imperfectives. In some LVCs nouns can be used both in singular and plural and/or in different cases. An important feature of LVCs is that they can frequently be substituted with a single "heavy" verb, e.g. *donijeti odluku – odlučiti*, although the meanings of the LVC and their paraphrases, i.e. semantically full verbs, often do not exactly correspond.

VP idioms (or *phrasemes*) are usually categorized into two groups: decomposable and non-decomposable idioms. The division is based on the degree of semantic and syntactic opaqueness of the whole construction in regard to its elements, as well as on the possibility of the word order change within an idiom.

The group of other verbal MWEs in our research refers to multiword predicates consisting of a finite verb and one or more verbs in infinitive form. Verbs in finite form typically belong to modal or phasal verbs (e.g. inchoative verbs).

All verbal MWEs listed above form complex sentence predicates, i.e. multiword units, and therefore need to be identified and annotated in Croatian language resources.

3. Procedure

There are three dependency treebanks available for Croatian. The first one is the Croatian Dependency Treebank (HOBS) that in its latest version encompasses 4,626 sentences of Croatian newspaper. HOBS is freely available for on-line search (hobs.ffzg.hr). The second one – SETIMES.HR dependency treebank (http://nlp.ffzg.hr/resources/corpora/setimes-hr/) – was built on top of the newspaper text from the SETIMES parallel corpus. The treebank contains approximately 9,000 sentences, and it is completely free. These two treebanks are annotated with modified versions of annotation schemes used in the Prague Dependency Treebank project done for Czech. However, we decided to deal with verbal MWEs in the third available treebank, Universal Dependency (UD) Treebank for Croatian.[1] The UD treebank of Croatian was also built from newspaper text originating from SETIMES parallel corpus, but annotated according to UD annotation. The UD treebank version used in this experiment consisted of 3557 sentences.

This treebank was chosen for the task presented in this paper for two reasons: 1) although it is the smallest in size compared to other two treebanks, it is large enough for a preliminary research of identification and annotation of verbal MWEs in Croatian, 2) the UD annotation guidelines account for different types of MWEs and mark the relation between their components on syntactic level. They distinguish between fixed multiword expressions (for example, *in spite of* is marked with *mwe* tag), multiword names (*name*) and foreign phrases (*foreign*). Other types of MWEs are recognized as well. For example, parts of English phrasal verbs are marked as compounds. However, the criteria for the recognition of MWEs are not clearly stated: "Deciding whether an expression in a language should be treated as a MWE is something that has to be decided for each language, and in some cases this will require somewhat arbitrary conventions, because it involves choosing a cut point along a path of

[1] A detailed account of building this treebank and achieved parsing scores is given in Agić and Ljubešić (2015).

grammaticalization."[2] For Croatian, this kind of convention is not established yet, and the following experiment is the first step in this direction.

In the first step of the task we built an initial list of verbal MWEs from available work done in this area for Croatian. A list of approximately 20 phrasal verbs (i.e. combinations consisting of a verb and a preposition) was taken from Katunar et al. (2012), whereas a list of LVCs was compiled from Silić and Pranjković (2005) and Gulić (2015). This list contained 80 LVCs for Croatian. We searched for verbal MWEs from this initial set in the chosen treebank. In this step we wanted to determine which MWEs appear in the treebank and whether they can be automatically annotated. We also wanted to determine whether the light verbs from the list can be used for the detection of other NPs or PPs in new LVCs. Unfortunately, the obtained results were completely unsatisfactory since none of the phrasal verbs was detected in the treebank whereas only 14 LVCs from the initial set were identified. This could suggest that the initial list is too small and narrow (and even not built on the real language data, i.e. data from various corpora) or that only a very limited number of verbal MWEs occurs in the corpus. The other option is not very likely since the corpus consists of newspaper texts and such constructions are very frequent in this genre. This was the reason to manually annotate the selected treebank for verbal MWE types as described in Section 2. In other words, in the second step of the task we manually annotated 3557 sentences from the UD treebank for Croatian for phrasal verbs, LVCs, VP idioms and multiword predicates. The results are presented and discussed in the following section.

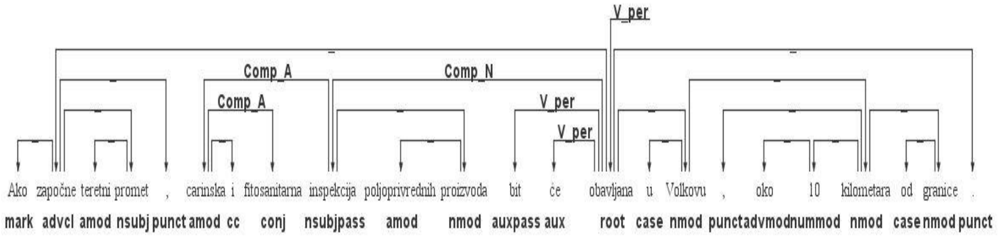

V_per
Comp_A
Comp_N
Comp_A
V_per
V_per
Ako započne teretni promet , carinska i fitosanitarna inspekcija poljoprivrednih proizvoda bit će obavljana u Volkovu , oko 10 kilometara od granice .
mark advcl amod nsubj punct amod cc conj nsubjpass amod nmod auxpass aux root case nmod punct advmod nummod nmod case nmod punct

Figure 1: An example of a sentence annotated for a LVC

Verbal MWEs were marked in the corpus on a separate level of annotation in order to enable their explicit differentiation from other verbal phrases with similar or identical morphological and syntactic properties. This is particularly important when dealing with LVCs and verbal idioms. Each member of verbal MWEs was marked in our approach. More details are given in the following section.

4. Results

The total number of verbal MWEs belonging to the group of phrasal verbs (cf. Section 2) is 371. We annotated verbs in such MWEs in the treebank with V_Prep tag and prepositions with Comp_Prep tag. In Table 1 we list the most frequent ten verbal MWEs annotated witih V_Prep tag in the treebank.

[2] UD annotiation guidelines, http://universaldependencies.org/u/overview/syntax.html

verbs	prepositions	frequency of MWEs (verb + preposition)	total frequency of verbs in the corpus (within and outside MWEs)
pozivati 'call'	*na* 'for'	18	30
razgovarati 'talk'	*o / s* 'about / to'	10	29
raditi 'work'	*na / o / za* 'on / for'	9	47
dovesti 'bring'	*do* 'to'	7	16
glasovati 'vote'	*za / o / protiv* 'for / on / against'	7	10
odnositi se 'refer'	*na* 'to'	7	17
ovisiti 'depend'	*o* 'on'	7	14
nastaviti continue'	*s* 'with'	6	32
sastati se 'meet'	*s* 'with'	6	37

Table 1: the most frequent 10 MWEs annotated as prepositional verbs and total frequency of verbs

The second group of verbal MWEs comprises LVCs. The total number of annotated LVCs in the treebank is 847. Verbal parts in these constructions are tagged with V_per tag. NPs and PPs that are elements of these constructions were annotated as Comp_N and Prep_N respectively. In Table 2 we present the most frequent ten light verbs in the selected treebank, NPs and PPs that co-occur in LVCs as well as their frequency in the corpus. The frequency threshold is set at two occurrences.

light verb	frequency in various LVCs	NPs in LVCs	frequency of NPs in LVCs	PPs in LVCs	frequency of PPs in LVCs
imati 'have'	73	*posljedice* 'consequences'	5	*u vidu* 'in sight' (keep in mind)	3
		pravo 'right' (be right)	5	*za cilj* 'as its aim'	3
		utjecaj 'influence'	5		
biti 'be'	56	*domaćin* 'host'	3	*u stanju* 'in position'	6
				u mogućnosti 'able to'	4

dobiti 'get'	30	*nagradu* 'prize' (receive a prize)	5	*na težini* 'on weight' (gain importance)	2
		potporu 'support'	5		
izraziti 'express'	25	*nadu* 'hope'	6		
		potporu 'support'	6		
		zabrinutost 'concern'	6		
osvojiti 'win'	18	*nagradu* 'an award'	7		
		odličje 'a medal'	3		
dati 'give'	17	*izjavu* 'statement' (make a statement)	2		
		potporu 'support'	2		
podnijeti 'submit'	16	*ostavku* 'resignation'	8		
		tužbu 'a complaint'	3		
postati 'become'	14	*članicom* 'member'	5		
predstavljati 'be'	13	*zapreku* 'obstacle'	4		
poduzeti 'take'	12	*korake* 'steps'	8		

Table 2: the most frequent 10 light verbs annotated as V_Per and their nominal components

The third group of verbal MWEs encompasses VP idioms. We have detected and marked 18 verbal MWEs as VP idioms. However, there are only 7 different VP idioms in the selected corpus. They are listed in Table 3, along with their overall frequency. Each member of verbal idioms was marked with V_idiom tag.

VP idiom	frequency
biti na čelu 'be at the head'	4
biti na klimavim nogama 'be without a firm foundation'	3
biti u punom zamahu 'be in full swing'	3
hvatati se / uhvatiti se u koštac 'take the bull by the horns'	3

ne odustati ni pedlja 'not to retreat a single inch'	2
staviti točku na 'put an end to'	2
zatražiti zeleno svjetlo 'request approval'	1

Table 3: 7 VP idioms and their overall frequency

Finally, in Table 4 we present the results for multiword predicates consisting of a verb in finite form and a verb in infinitive form. The elements of these MWEs are marked as V_fin and V_inf respectively.

verb in finite form	frequency
moći 'can'	142
trebati 'should'	120
morati 'must'	95
željeti 'want'	32
biti 'be'	14
planirati 'plan'	14
pokušati 'try$_{pf}$'	12
pokušavati 'try$_{ipf}$'	10
odlučiti 'decide'	9
kaniti 'plan'	9
uspjeti 'succeed'	8
nastaviti 'continue'	7
početi 'begin'	5
htjeti 'will'	5
odbiti 'refuse'	5

Table 4: the most frequent 15 verbs annotated as V_fin and their overall frequency

5. Discussion

The first group contains 371 verbs that form so called phrasal or prepositional verbs in Croatian. PPs in this group should be differentiated from PPs that denote adverbials. The PPs in this group denote objects. Semantically similar prepositional objects can be introduced with different prepositions. In some cases the meaning of the verb is not affected by the selection of a preposition, e.g. *misliti na* 'to think of' and *misliti o* 'to think about'. In other cases the meaning of the verb alters under the influence of the preposition introducing the object, e.g. *odnositi se na* 'to refer to' and *odnositi se prema* 'to treat somebody in a particular way'. In future work these information will be used for the creation of verb valency frames and distinguishing of senses in the large database of Croatian verbs CroDeriV.[3]

The second group encompasses light verb constructions with 252 unique light verbs. The light verbs from these semi-compositional constructions always have their counterparts that are not impoverished in their lexical meaning. The light verbs retain only a portion of the full lexical meaning of their homonymic counterparts. As the obtained results reveal, this group can be further divided into several subgroups. The first division is based on the ability to be paraphrased with a single verbs (e.g. *dati doprinos* 'to give a contribution' – *doprinijeti* 'to contribute'). However, in numerous cases such paraphrases are not possible, e.g. *dobiti zadatak* 'to get an assignment'. On top of that, some LVCs that can be paraphrased with a single verb in certain contexts, in other contexts acquire additional semantic components and paraphrases are not possible. E.g. the construction *donijeti zaključak* 'to make a conclusion' can be paraphrased with the verb *zaključiti* 'to conclude'. They are not completely interchangeable in all contexts, since the LVC *donijeti zaključak* can in some contexts mean 'to agree that'. This LVC can in some cases imply that the conclusion(s) are presented or given in a written form, whereas the verb *zaključiti* almost never appears in this context. The group of detected LVC is, as far as we know, the biggest list of such constructions available for Croatian.[4]

The third group contains VP idioms. For several reasons this is the most problematic group. Firstly, the inter-annotator agreement was extremely low when dealing with this category. There was a significant overlapping of VP idioms and LVCs. Secondly, the lack of clear criteria for distinguishing LVCs and VP idioms in Croatian literature made the whole procedure even more complicated. Finally, the results show that the division of VP idioms into decomposable and non-decomposable VP idioms discussed in Section 2 seems to have no relevance for such constructions in Croatian since all detected and annotated VP idioms belong to the group of decomposable VP idioms.

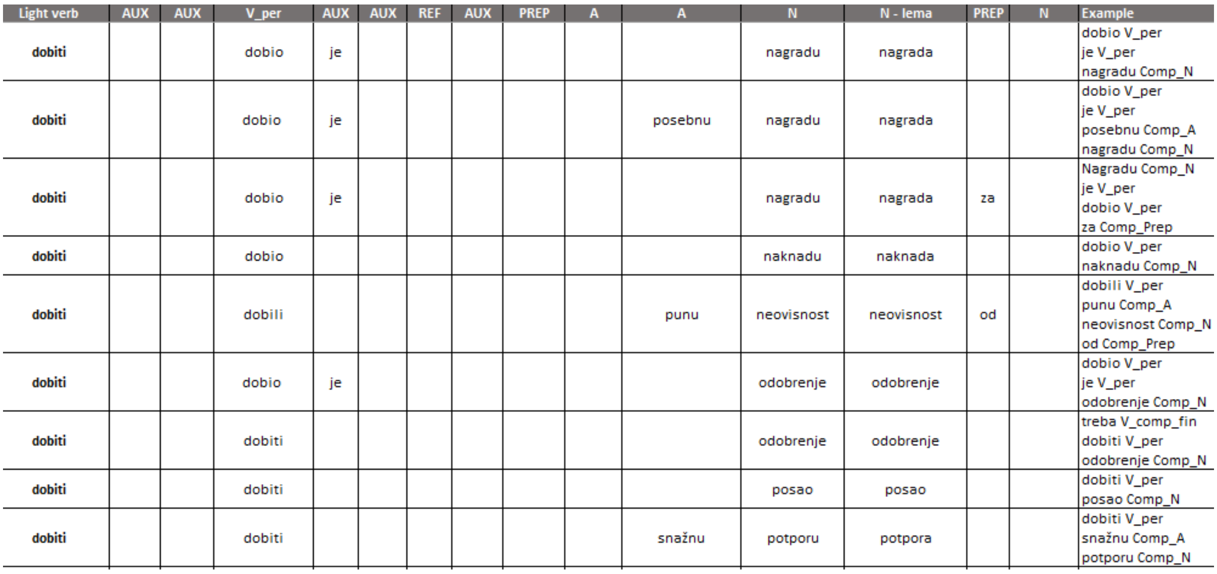
The fourth group contains multiword predicates consisting of a verb in finite form followed by one or more verbs in infinitive forms. Verbs that appear in finite forms predominantly belong to modal verbs (e.g. must, should etc.) or phasal verbs (begin, start etc.) However, other detected verbs are those that are usually not classified as modal or phasal in Croatian (e.g. *planirati* 'to plan', *pokušati* 'to try', *uspjeti* 'to succeed', *odlučiti* 'to decide' etc.). These results address the issue of redefinition verbal groups that are followed by infinitive forms as well as the treatment of such constructions as complex predicates. In numerous cases infinitive VPs appear to be morphosyntactic realization of objects. Finally, infinitives often follow nominal predicates (e.g. in constructions as *biti voljan učiniti* 'to be willing to do') or LVCs (e.g. *biti u mogućnosti doći* 'be able to come'). These constructions raise additional questions regarding the status of complex predicates and the traditional notion of object. However, this topic is beyond the scope of this contribution.

6. Conclusion and Future Work

It is clear that the results obtained for all four verbal MWEs in Croatian are valuable in several respects: They were obtained from the first research of such constructions for Croatian that is based on corpus data and therefore more truly indicate the productivity of particular prepositional and light verbs in combinations with various PPs and NPs than the data presented in existing literature. Secondly, the results enable further investigation of possibilities for automatic detection and recognition of MWEs both from monolingual and parallel corpora of the Croatian language. Thirdly, the presented results raise several theoretical questions and provide possibilities for their in-depth analysis. Finally, the obtained results will enable the creation of a language resource that would encompass various types of verbal MWEs and enable queries according to various parameters. The outline of this database is given in Figure 2 below.

[3] CroDeriV in its present shape contains data on derivational relatedness of Croatian verbs. It is available at http://croderiv.ffzg.hr/. The next phase of the development is aimed at valency and meaning description of verbs.

[4] All the results discussed here are available upon request. The complete database will be public and downloadable.



Light verb	AUX	AUX	V_per	AUX	AUX	REF	AUX	PREP	A	A	N	N - lema	PREP	N	Example
dobiti			dobio	je							nagradu	nagrada			dobio V_per je V_per nagradu Comp_N
dobiti			dobio	je						posebnu	nagradu	nagrada			dobio V_per je V_per posebnu Comp_A nagradu Comp_N
dobiti			dobio	je							nagradu	nagrada	za		Nagradu Comp_N je V_per dobio V_per za Comp_Prep
dobiti			dobio								naknadu	naknada			dobio V_per naknadu Comp_N
dobiti			dobili							punu	neovisnost	neovisnost	od		dobili V_per punu Comp_A neovisnost Comp_N od Comp_Prep
dobiti			dobio	je							odobrenje	odobrenje			dobio V_per je V_per odobrenje Comp_N
dobiti			dobiti								odobrenje	odobrenje			treba V_comp_fin dobiti V_per odobrenje Comp_N
dobiti			dobiti								posao	posao			dobiti V_per posao Comp_N
dobiti			dobiti							snažnu	potporu	potpora			dobiti V_per snažnu Comp_A potporu Comp_N

Figure 2: An excerpt from the database of verbal MWEs

Acknowledgement

This paper is partially supported by European Union, European Social Fund, under the project grant HR.3.2.01.-0037 Mrežni portal za online učenje hrvatskoga jezika HR4EU.

References

Agić, Željko and Ljubešić, Nikola (2015). Universal Dependencies for Croatian (that Work for Serbian, too). Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pp. 1–8, Hissar, Bulgaria

Baldwin, Timothy and Su Nam Kim (2010). Multiword Expressions. In: Nitin Indurkhya and Damerau, Fred J. (Eds.) *Handbook of Natural Language Processing, Second Edition*. Boca Raton, USA: CRC Press, pp. 267-292.

Gulić, Anamarija. (2015). *Klasifikacija perifraznih glagola u hrvatskom jeziku.* MA thesis. Faculty of Humanities and Social Sciences, University of Zagreb.

Katunar, D., Srebačić, M., Raffaelli, I., and Šojat, K. (2012). Arguments for Phrasal Verbs in Croatian and Their Influence on Semantic Relations in Croatian WordNet. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Available at: https://bib.irb.hr/datoteka/582794.Croatian_phrasal_verbs_KSRS.pdf

Rosén, V. et al. (2014). A Survey of Multiword Expressions in Treebanks. In: *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT14)*, 11–12 December 2015, Warsaw, Poland.

Rosén, V. et al. (2016). MWEs in Treebanks: From Survey to Guidelines. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16),* 23-28 May 2016, Portorož, Slovenia

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, Mexico, pp. 1-15.

Silić, J. and Pranjković, I. (2005). *Gramatika hrvatskoga jezika – za gimnazije i visoka učilišta.* Zagreb: Školska knjiga.