Verbal multiword expressions in the Croatian verb valency database

Ivana Brač **Lobel Filipić** **Maja Matijević** **Siniša Runjaić**
Institute for the Croatian Language
ibrac@ihjj.hr lfilipic@ihjj.hr mmatijevic@ihjj.hr srunjaic@ihjj.hr

Relevant UniDive working groups: WG2, WG1

1 Introduction

Multiword expressions in running text pose challenges in NLP (Constat and Nivre, 2016; Savary et al., 2017), as well as in lexicographic resources (Koeva et al., 2016) and theoretical syntactic and semantic research. In this poster, we will present the treatment of reflexive verbs, verbal idioms, and light verb constructions in Croatian general language online dictionaries (Croatian Language Portal; Croatian Web Dictionary - Mrežnik) and online valency lexicons (CROVALLEX; e-Glava), with a focus on proposing solutions for the description of verbal multiword expressions in the verb valency database Verbion.

2 Description of reflexive verbs and verbal idioms

Regarding reflexive verbs, if they are inherently reflexive, in all analysed resources, they are listed as lemmas (*nadati se* 'hope'), while verbs with transitive and reflexive variants have different treatments. In Mrežnik (Hudeček and Mihaljević, 2020), the clitic *se* is written in brackets next to a verb as lemma (e.g., *brijati (se)* 'shave (oneself)') or as sublemma (see *lupati se* 'hit oneself'). In HJP, the reflexive variant is one of the verb senses, as well as in e-Glava (Birtić et al., 2017). In CROVALLEX (Preradović, 2020), they are lemma (*buditi se* 'wake up') or it is not emphasized that a verb is reflexive (see *tuširati se* 'take a shower'). In our database, the reflexive variant will be listed as a sublemma. In general language dictionaries, verbal idioms are a separate category within lexicographic entry. It has to be emphasized that in Mrežnik, an explanation of the verbal idiom and an example from the corpora are provided. Regarding valency lexicons, in CROVALLEX, verbal idioms are a separate verb meaning, marked as an idiom, while in e-Glava, the verbal idiom, its explanation and an example from the corpora are a separate category related to the verb lemma. In our database, the solution from e-Glava will be adopted (a).

a) **dobiti svoje**

'to get one's = to get what one deserves; to face the music'
biti kažnjen, proći loše vlastitom krivicom
'to suffer the consequences'

(1) *Još će dobiti svoje kad dozna Milan.*
yet will.3sg get.inf one's.acc.sg when findout.prs.3sg Milan.nom.sg
'One will get one's due when Milan finds out.'

3 Light verb constructions

Dealing with light verb construction is more complex. In certain constructions, it is sometimes difficult to determine whether it is a semantically full (main) verb, semantically empty or semantically bleached. There is also a question of semantic roles assignment; whether the semantic role is assigned by the verb, the noun, or by both (Grimshaw and Mester, 1988; Butt, 2010; Wittenberg, 2014). Determining the predicative noun in LVC also varies, as some approaches define it as a direct object or as a part of the predicate. Regarding general language online dictionaries, LVCs are listed as a separate sense of the verbs only in Mrežnik. That sense includes a generic definition: 'VERB appears as a light verb with nouns and can often be replaced by a full verb related to the corresponding noun', along with single-word synonym verbs and examples from the corpora. In our database, the three-level description from e-Glava (Birtić et al., 2017) is largely retained. The first level contains verb lemmas, semantic classes, a morphological block containing inflectional forms, along with the category of verbal idioms with their explanations and examples from the corpora. The second level encompasses verb senses with corresponding definitions. The third level consists of morphological, syntactic, and semantic descriptions of dependents, as shown in Table 1 - Table 3. For each participant of an event, semantic role, syntactic phrase, its morphological realization, and lexical units from

the example are provided, as well as the most frequent lexemes from the Croatian web corpus – hrWaC (Ljubešić and Klubička, 2014). The examples are annotated using UDPipe (Straka et al., 2016). In (b), the processing of the semantically full verb *napraviti* 'make' is presented, while (c) and (d) display the light verb. The light verb is listed as a separate meaning, following conclusions that light verbs have meaning, not only function (Butt, 2010; Jackendoff, 2007), and it is accompanied by a predicative noun that gives it full meaning. In (c), the description of the light verb construction *napraviti (po)grešku* ('make a mistake') is shown. The only lexeme that may be lexicalized in the object position is *(po)greška* 'mistake'. A more complex example is (d) since the verb selects a larger class of nouns, such as *analiza* 'analysis', *procjena* 'evaluation', *provjera* 'check' and *istraživanje* 'research'. It is decided to write it as shown in Table 3 because these VMWEs represent closely related meanings. The question is if the nouns *pogreška* ('mistake'), *analiza* 'analysis', *procjena* 'evaluation', etc. bear a semantic role. The answer could be that the direct object NPs in LVCs do not bear a semantic role since the light verb cannot assign it, but due to the argument transfer, the direct object NPs transfer their argument structure to the argument structure of the light verb and as a result, the semantic role of the Theme is assigned to the indirect object in the genitive case (Grimshaw and Mester, 1988; Karimi-Doostan, 2005). Another answer could be that due to the argument sharing, both the light verb and the noun assign semantic roles (Culicover and Jackendoff, 2005; Jackendoff, 2007; Butt, 2010). At the moment, we decided to follow the latest conclusion and define the semantic role of the direct object as 0 and Theme in Table 2 and Table 3.

b) **napraviti**
'make'
stvoriti što, dati čemu oblik
'create something, give shape to something'

(2) *Teta, [ti] napravi nam*
aunt.voc.sg you.nom.sg make.imp we.dat.pl
veliku kuću.
big.acc.sg house.acc.ag
'Aunt, make us a big house.'
See Table 1.

(c) **napraviti grešku/pogrešku**
'make a mistake'
postupiti netočno ili neispravno u kakvu postupku; pogriješiti
'to proceed inaccurately or incorrectly in a procedure; make a mistake'

(3) *Osim toga napravili smo*
besides that.gen.sg make.pst.pl AUX
jednu ogromnu grešku.
one.acc.sg huge.acc.sg mistake.acc.sg
'Besides that, we made a huge mistake.'
See Table 2.

(d) **napraviti analizu / procjenu / provjeru / istraživanje**
'conduct an analysis/evaluation/check/research'
proučiti što kako bi se utvrdilo kakvo stanje; analizirati; procijeniti; provjeriti; istražiti
'examine something to determine its current condition; analyse; evaluate; check; research'

(4) *Banke moraju napraviti*
bank.nom.pl must.prs.3pl conduct.inf
odgovarajuću analizu rizika
proper.acc.sg analysis.acc.sg risk.acc.sg
projekta.
project.gen.sg
'Banks must conduct a proper risk analysis of the project.'
See Table 3.

Since this is a work in progress, and we are at the very beginning of the project, our goal is to gather feedback on solutions in our database to improve it and possibly align it more with other resources.

4 Acknowledgements

This research has been conducted within the project *Semantic-Syntactic Classification of Croatian Verbs* (IP-2022-10-8074), supported by the Croatian Science Foundation, and NextGeneration EU.

References

Matea Birtić, Ivana Brač, and Siniša Runjaić. 2017. The main features of the e-glava online valency dictionary. In *Proceedings of eLex 2017 conference*, pages 43–46.

Miriam Butt. 2010. The light verb jungle: Still hacking away. In *Complex predicates in cross-linguistic perspective*, pages 48–78.

Matthieu Constat and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic

analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 161–171.

Peter W. Culicover and Ray Jackendoff. 2005. *Simpler Syntax*. Oxford University Press, Oxford.

Jane Grimshaw and Armin Mester. 1988. Light verbs and -marking. *Linguistic Inquiry*, 19/2:205–232.

Lana Hudeček and Milica Mihaljević. 2020. The croatian web dictionary – mrežnik project – goals and achievements. *Rasprave*, 46/2:645–667.

Ray Jackendoff. 2007. A parallel architecture perspective on language processing. *Brain research*, 1146:2–22.

Gholamhossein Karimi-Doostan. 2005. Light verbs and structural case. *Lingua*, 115:17372–1756.

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, and Svetlozara Leseva. 2016. Semi-automatic compilation of the dictionary of bulgarian multiword expressions. In *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology, Workshop at LREC2016*, pages 86–95.

Nikola Ljubešić and Filip Klubička. 2014. bs,hr,srwac - web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35.

Nives Mikelić Preradović. 2020. *CROVALLEX: valencijski leksikon glagola hrvatskoga jezika*. Filozofski fakultet, Zagreb.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions. Association for Computational Linguistics*, pages 31–47.

Milan Straka, Jan Hajič, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Eva Wittenberg. 2014. *With Light Verb Constructions from Syntax to Concepts*. Potsdam Cognitive Science Series, Potsdam.

	d1	d2	d3
Semantic role	Agent	Theme	Benefactive
Syntactic phrase	NP	NP	NP
Morphological realization	nom	acc	dat
Lexemes	teta ‘aunt’	kuću ‘house’	nam ‘us’

Table 1: Description of the verb *napraviti* ‘make’

	d1	d2
Semantic role	Agent	0/Theme
Syntactic phrase	NP	NP
Morphological realization	nom	acc
Lexemes	ja ‘I’	grešku ‘mistake’
		pogrešku ‘mistake’

Table 2: Description of the LVC *napraviti (po)grešku* ‘make a mistake’

	d1	d2	d3
Semantic role	Agent	0/Theme	Theme
Syntactic phrase	NP	NP	NP
Morphological realization	nom	acc	gen
Lexemes	banke ‘banks’	analizu ‘analysis’	učinaka ‘performance’
		procjenu ‘evaluation’	stanja ‘condition’
		provjeru ‘check’	uzorka ‘sample’
		istraživanje ‘research’	tla ‘soil’
			tržišta ‘market’
			rizika ‘risk’
			poslovanja ‘business’
			troškova ‘costs’

Table 3: Description of the LVC *napraviti analizu/procjenu/provjeru/istraživanje* ‘conduct an analysis/evaluation/check/research’