Verbal Multiword Expressions in the Croatian Verb Lexicon

Ivana Brač
Institute for the Croatian language
ibrac@ihjj.hr

Matea Birtić
Institute for the Croatian language
mbirtic@ihjj.hr

Abstract

The paper examines the complexities of encoding verbal multiword expressions in the Croatian verb lexicon. The lexicon incorporates a verb's description at the syntactic, morphological, and semantic levels. This study explores the treatment of reflexive verbs, light verb constructions, and verbal idioms across several Croatian and Slavic language resources to find the best solution for the verb lexicon. It addresses the following research questions: 1. How should reflexive verbs, i.e., verbs with the reflexive marker *se*, be treated? Should they be considered as separate lemmas, sublemmas of non-reflexive counterparts, or as one of their senses? 2. What syntactic label and semantic role should be assigned to a predicative noun in light verb constructions? 3. Should verbal idioms be included, and, if so, at which level of a description? Our conclusion is that all reflexive verbs should be treated as separate lemmas since they are distinct lexemes that have undergone semantic and syntactic change. To differentiate between a semantically full verb and a light verb, we have introduced the label LV and decided not to assign a semantic role to a predicative noun. By including verbal idioms and their translation into English, non-native users can benefit from the lexicon. The aim is to enhance the verb lexicon for the more effective description and recognition of verbal multiword expressions.

Keywords: reflexive verbs, light verbs constructions, verbal idioms, verb lexicon.

1 Introduction

Multiword expressions (MWEs) in running text pose challenges in natural language processing (e.g., Sag et al., 2002; Constant and Nivre, 2016; Savary et al., 2017; Osenova and Simov, 2018), in lexicographic resources (Koeva et al., 2016), and in theoretical syntactic and semantic research (e.g., Grimshaw and Mester, 1988; Butt, 2010). A verb lexicon containing the description of verbal multiword expressions (VMWEs) (see Kettnerová and Lopatková, 2013), such as the Croatian verb lexicon Verbion, can be useful for improving the accuracy and efficiency of processing and in understanding these expressions in various linguistic applications.

The verb lexicon Verbion contains a verb lemma accompanied by labels about aspect and reflexive uses, a morphological block with inflectional forms, and verb senses. Each sense is accompanied by a semantic class, a semantic frame from FrameNet (Baker et al., 2003), and one or more valency frames. Each valency frame includes an example from corpora which is analyzed at syntactic, morphological, and semantic levels. For the syntactic description, phrase type labels such as NP, PP, etc., are used, and for the semantic level, semantic roles mainly adopted from VerbNet (Kipper Schuler, 2005) and frame elements from FrameNet (Baker et al, 2003) are assigned. In the first phase of the project, a description of the 500 most frequent verbs in Croatian will be provided, and the results will be publicly available on the project's webpage (https://semtactic.jezik.hr/). The questions that need to be answered at this stage of planning are: 1. How should reflexive verbs, i.e., verbs with the reflexive marker *se*, be treated? Should they be considered as separate lemmas, sublemmas of transitive or (rarely) intransitive verbs, or as one of the senses of their non-reflexive counterparts? 2. What syntactic label and semantic role should be assigned to a predicative noun in light verb constructions (LVCs)? 3. Should verbal idioms be included, and, if so, at which level of a description?

In this paper, we present the treatment of reflexive verbs, light verb constructions, and verbal idioms in Croatian general language online dictionaries (*Hrvatski jezični portal* = Croatian Language Portal: `https://hjp.znanje.hr/`; the online version of *Hrvatski školski rječnik* = Croatian School Dictionary: `https://rjecnik.hr/`; *Hrvatski mrežni rječnik* = Croatian Web Dictionary – Mrežnik: `https://rjecnik.hr/mreznik/`) and online valency lexicons (CROVALLEX: `http://theta.ffzg.hr/crovallex/`; e-Glava: `http://valencije.ihjj.hr`), as well as other Slavic valency lexicons with rich syntactic and semantic descriptions, such as the Czech VALLEX (Kettnerová and Lopatková, 2013, Kettnerová, 2023). In Section 2, the treatment of inherently reflexive verbs or reflexiva tantum, "proper" reflexive, derived reflexive, and reciprocal reflexive verbs is presented. Section 3 addresses the processing of light verb constructions. Section 4 presents the recording of verbal idioms, which is followed by the Conclusion. Our aim is to present solutions for the description of VMWEs at syntactic and semantic levels in the Croatian verb lexicon Verbion.

2 Reflexive verbs

There are many classifications of reflexive verbs made both for the Croatian language and crosslinguistically. As for Croatian, at one end are 1. inherently reflexive verbs or *reflexiva tantum*, which cannot appear without the reflexive marker *se* (e.g., *smijati se* 'laugh', *natjecati se* 'compete'), and at the other end there are 2. "proper" (or true) reflexive verbs, which are basically transitive verbs whose object can be replaced by the reflexive pronoun *sebe* 'oneself' or its shorter variant *se* (e.g., *češljati se(be)* 'comb oneself').[1] There is a third distinct group: 3. reciprocal reflexive verbs (e.g., *ljubiti se* 'kiss each other'). We could add a fourth group – 4. derived reflexives – a group that is between inherently reflexive verbs and "proper" reflexive verbs. They have transitive and intransitive counterparts, but the reflexive marker *se* cannot be replaced by the pronoun *sebe* (*igrati* 'play (tran.)' – *igrati se* 'play (refl.)'.

Since there are many yet unaligned approaches, we investigated the treatment of these four groups of verbs in the resources mentioned in the Introduction. *Mrežnik* (Hudeček and Mihaljević, 2020) and *Hrvatski školski rječnik* (ŠR) (Birtić et al., 2012) list inherently reflexive verbs as separate lemmas (e.g., *čuditi se* 'wonder'). "Proper" or syntactically reflexive verbs are also listed with the reflexive marker, but it is placed in parentheses (e.g., *kupati (se)* 'bathe (oneself)'). If the meaning of a transitive verb and its reflexive variant which belongs to "proper" reflexive verbs is the same with a different object reference, the reflexive marker *se* in parentheses is written next to the headword. For example, *kupati (se)* is the main lemma which means 'to wash somebody in the bathtub or container full of water'. If there is another meaning of the reflexive variant, it is written as an additional sense and the reflexive verb is repeated as a sublemma without parentheses: *kupati se* 'be in water or swim'. If the meanings of transitive and reflexive verbs are not similar, the reflexive variant is listed only as one of the senses of the transitive verb. For example, in both *Mrežnik* and ŠR the verb *češljati (se)* 'comb (refl.)' is recorded only as one sense of the lemma *češljati* 'comb', and in this case, *se* is placed in parentheses. The derived reflexive verb *buditi se* 'wake up' is treated parallel to the "proper" reflexive verb *kupati se* 'bathe': the reflexive marker *se* is placed in parentheses next to the main lemma and is also listed as a separate sublemma under one of the main verb's senses. Reciprocal reflexive verbs, which are not reflexiva tantum, are treated as a sense of a transitive verb. For example, *dogovarati se* 'arrange things together' is a sense of the lemma *dogovarati* 'arrange, fix'.

Hrvatski jezični portal (HJP), like all other Croatian dictionaries, treats inherently reflexive verbs as separate lemmas. However, the treatment of other groups is highly inconsistent. For example, the verbs *kupati* 'bathe' and *kupati se* 'bathe oneself' are listed as two separate headwords. In contrast, the verb *češljati se* 'comb oneself' is treated as a sublemma of the main lemma *češljati* 'comb' and is recorded as the fourth sense of the

[1] There are many discussions regarding the status of *se* in Croatian. Some authors consider it a pronoun (Barić et al., 1997; Raguž, 2010), others view it as a particle (Babić et al., 2007; Oraić Rabušić, 2018), while some argue that it is a particle with reflexiva tantum and derived reflexive verbs, and a pronoun with "proper" reflexive verbs (Silić and Pranjković, 2005; Belaj, 2001). In a recent work, Belaj (2024) distinguishes between single-participant and multiparticipant middle verbs, i.e., reciprocal middle verbs. In both cases, Belaj (2004, 98) considers *se* as an integral part of the verb, and consequently a particle. We believe that *se* diachronically originates from the pronoun *sebe*, but, synchronically, it is a particle.

headword *češljati* 'comb'. The verb *prati se* 'wash oneself' is treated differently from both *kupati se* 'bathe oneself' and *češljati se* 'comb oneself'. The reflexive marker *se* is not introduced in the morphological block and there is no separate sense for *prati se* 'wash oneself', but the reflexive marker *se* is listed next to the object pronoun within the definition of the transitive verb *prati* 'wash', see (1).

(1) *prati* 'wash'
1. *(koga, što, se) ispiranjem u tekućini (ob. u vodi) uklanjati nečistoću*
'(somebody, something, oneself) remove dirt by rinsing in a liquid (usually in water)'

In the morphological block of the derived reflexive verb *buditi se* 'wake up', *se* is recorded next to the object pronoun *što* 'something', but both senses (transitive and reflexive) are defined together. The treatment of reciprocal reflexive verbs in HJP does not follow any uniform pattern either. For example, *dogovarati se* 'arrange things together' is a separate lemma since the HJP does not contain the verb *dogovarati* 'arrange, fix', and the reciprocal usage of the verb *sresti se* 'meet each other' is not recorded at all. With the verb *ljubiti se* 'kiss each other; to love each other', the reflexive marker *se* is listed in the morphological block, next to the object pronoun *koga, što* 'somebody, something'. This is very confusing since there is no indication of to which of the listed senses the reflexive marker is connected. If *se* is a marker of the "proper" reflexive verb, it means 'love oneself' and if the reflexive marker *se* denotes reciprocity, the verb means 'kiss one another or to love one another'. It seems that reciprocity is marked in HJP only if the verb is introduced as a separate reflexive verb, and its definition indicates that the meaning is reciprocal.

In CROVALLEX (Mikelić Preradović, 2020), reflexiva tantum, derived, and reciprocal reflexive verbs are recorded as separate lemmas. Therefore, for example, the verbs *penjati se* 'climb' (reflexiva tantum), *buditi se* 'wake up' (which is considered a derived reflexive verb) and *ljubiti se* 'kiss each other; love each other' (as a reciprocal verb) are introduced as separate lemmas. In contrast, *prati se* 'to wash oneself' (a "proper" reflexive verb) is treated as one of the senses of the verb *prati* 'wash'. However, the sublemma is not accompanied by the marker *se* nor is there any label indicating reflexivity (2). Reflexivity is only visible in the example and in the verb's definition, which contains the reflexive pronoun *sebe* 'oneself'. This could pose a problem for non-native users of the dictionary.

(2) 4. *prati (pràti) ≈ uklanjati sa sebe prljavštinu vodom i sapunom*
'wash ≈ to remove dirt from oneself with water and soap'
-example: *Kupač se prao četkom* 'The swimmer washed himself with a brush.'
-class: dress

In e-Glava (Birtić, Brač, and Runjaić, 2017), which contains only approximately 50 psychological verbs, the main principle for handling reflexive verbs is to treat reflexiva tantum as a separate lemma exclusively. All other reflexive verbs, including those with transitive and intransitive counterparts, as well as reciprocals, are not listed as separate lemmas but rather as senses of the main lemma. Additionally, they are accompanied by the label *pov.* 'refl.' to indicate reflexivity. If a verb can have both derived reflexive and reciprocal reflexive variants, these are introduced as separate sublemmas, each with its distinct definition. In (3), the second sense pertains to the derived reflexive verb, whereas the fourth is the reciprocal reflexive verb. Among psychological verbs, there are no examples of "proper" reflexive verbs in e-Glava; however, it is presumed they would be treated similarly to derived reflexive verbs.

(3) 1 *vrijeđati* 'to insult' *nanositi uvrede komu, često riječima ili postupcima* 'to inflict insults on someone, often through words or actions'
2 *vrijeđati se* 'to take offense' *povr.* 'refl.' *osjećati se uvrijeđen, često čime; primati uvrede* 'to feel offended, often by something; to receive insults'
3 *vrijeđati* 'to irritate' *pobuđivati bol nadražujući bolno mjesto* 'to provoke pain by irritating a sore spot'
4 *vrijeđati se* 'to insult each other' *povr.* 'refl.' *nanositi uvrede jedan drugomu* 'to inflict insults upon each other'

The Czech VALLEX is available in several versions. In VALLEX 4.0 and 4.5, which include a data component, i.e., valency frames for active, non-reflexive and non-reciprocal uses of verbs, and a grammar component, i.e., derived valency frames, such as passive, reflexive and reciprocal uses of verbs (see Kettnerová et al. 2022: 40), a distinction is made between reflexiva tantum and derived reflexive verbs. Each verb classified as reflexiva tantum is assigned the attribute “reflexverb” with the value “tantum”, while derived reflexive verbs are assigned the attribute “reflexverb” with the value “derived”. Additionally, derived reflexive verbs are categorized into seven groups: decausative, autocausative, ‘partitive object’, reciprocal, converse, quasiconverse, and deaccusative. All reflexives are treated as separate lemmas.

In Verbion, all verbs with the reflexive marker *se* are considered as separate lexemes; therefore, they are separate lemmas. The reflexiva tantum is labeled with “REFL tantum” (e.g., *nadati se* ‘hope’) (4).

(4) **NADATI SE**
Eng. *hope*
1. *očekivati da će se ostvariti željeno*
‘to expect that the desired will be fullfilled’
Semantic class: Psych-Verbs
Frame: Desiring
Example: *Iskreno, **nadam se** čudu.*
‘Sincerely, I hope for a miracle.’

EX:	(ja)	**se nadam**	čudu
SYN:	NP	V REFL tantum	NP
MORPH:	NOM		DAT
SemR:	Experiencer		Stimulus
FE:	Experiencer		Event

Other reflexives are also considered as separate lemmas, as they have undergone semantic and syntactic changes and *se* is not a verb argument but a particle due to desemanticization, i.e., the loss of the semantic function (see Belaj, 2024: 100). The meaning of the reflexive verbs may be predictable from an (in)transitive use, e.g., the aforementioned verb *prati* ‘wash’ and *prati se* ‘wash oneself’. However, some reflexive verbs show only an indirect semantic and syntactic relation to the (in)transitive verbs (see Kettnerová et al. 2022: 45). For example, the transitive verb *praviti* ‘make, create’ means ‘to act with the intention of creating something; produce, create’ and it has a direct object in the accusative case as a complement, while the reflexive verb *praviti se* means ‘to attribute to oneself qualities that are not real; pretend’ with the predicative complement realized as NP or AP in the nominative or instrumental cases. Additionally, they belong to different semantic classes; the transitive verb belongs to the class of verbs of creation and transformation, while the reflexive verb belongs to the class of verbs with predicative components (see in Levin, 1993).

In Verbion, derived reflexive verbs are labeled with “REFL derived” (5), and reciprocal reflexive verbs with “REFL recipr” (6).

(5) **BUDITI SE**
Eng. *wake up*
1. *prestajati spavati, dovoditi se u budno stanje*
‘to stop sleeping, to bring oneself into a wakeful state’
Semantic class: Verbs of Change of State
Frame: Waking_up
Example: ***Budi** li **se** dijete zbog zubića…*
‘If the child is waking up because of teething…’

EX:	dijete	**se budi**	zbog zubića
SYN:	NP	V REFL derived	PP
MORPH:	NOM		zbog + GEN
SemR:	Agent_involun./Experie.		Cause
FE:	Sleeper		Explanation

(6) **LJUBITI SE**
Eng. *kiss each other*
1. *uzajamno izmjenjivati poljupce*
‘to mutually exchange kisses’
Semantic class: Verbs of Contact
Frame: Manipulation
Example: *Zatim **se** par strastveno **ljubio**.*
‘Then, the couple was kissing passionately.’

EX:	par	**se ljubio**	strastveno
SYN:	NP	V REFL recipr	AdvP
MORPH:	NOM		adv
SemR:	Agent		Manner
FE:			

Example: *Ona* ***se*** *strastveno* ***ljubi*** *s novim dečkom*.
'She is passionately kissing (with) her new boyfriend.'

EX:	Ona	**se ljubi**	s novim dečkom
SYN:	NP	V REFL recipr	PP
MORPH:	NOM		s + INST
SemR:	Agent		Co_Agent
FE:			

By treating reflexive verbs as separate lemmas and introducing the aforementioned labels, we believe that other resources dealing with reflexive verbs could benefit from more accessible and precise data. Since this resource is a dynamic database, it offers the flexibility to record reflexive verbs as separate lemmas, and ultimately enhancing the usability and accessibility of the resource for researchers dealing with the topic.

3 Light verb constructions

Dealing with light verb constructions is significantly more complex from syntactic, semantic, and technical perspectives. In certain constructions, it can be challenging to distinguish between a semantically full (main) verb and a semantically bleached verb.[2] Consequently, determining whether NP functions as an object or as part of the predicate can be difficult. There is also the question of semantic roles assignment, whether the verb assigns the role itself, the predicative noun, or both (see Grimshaw and Mester, 1988; Butt, 2010; Wittenberg, 2014). Since a light verb and a predicative noun form a single unit, it needs to be decided how to show it in the database.

In order to find the best solution for the database, we consulted other resources that included descriptions of LVCs. In the Czech VALLEX, a light verb determines a syntactic structure, i.e., valency frames, which are identical to those of their full verb counterparts. However, a predicative noun provides semantic participants since the verb is semantically incomplete (Kettnerová and Lopatková, 2013) via coreference (Alonso Ramos, 2007). A predicative noun in LVCs is labeled as Compound PHRase (CPHR functor) (Kettnerová et al., 2018), in contrast to a full verb valency frame where a noun in that position is labeled as Patient. It is neither an actant nor a free modifier, i.e., it is not a participant, and thus, a semantic role is not assigned to it (Cinková and Kolářová, 2006; Kettnerová and Lopatková, 2013). However, according to Kettnerová and Lopatková (2013), the number of complements in LVCs can be reduced in comparison with the full verbs since only verb complements that are semantically linked to the noun complements can be realized. Noun complements that are not linked to the verb complements remain on the surface structure as noun complements. An exception is causative light verbs with an Instigator or Causator in the subject position (Kettnerova et al., 2018), which is assigned by the verb, while other complements are semantic actants of the noun. In VALLEX 3.5, 4.0 and 4.5, LVCs are listed as senses "complex predicates (light verb)", accompanied by a frame, light verb constructions that contain nouns belonging to the same or a similar semantic class and that form a complex predicate with an LV, and a map. A verb can have more senses defined as complex predicates, depending on the variety of different frames (see, e.g., *činit*, *činívat*). Each noun has its own frame, which can be accessed by clicking on the noun.[3]

Regarding Croatian resources, in CROVALLEX (Mikelić Preradović, 2020), there is no detailed description of LVCs since it is based on VALLEX 1.0. LVCs are listed as verb senses along with an equivalent simple verb, but without defining the frame and the semantic class of the verb. They are marked as idioms, without distinguishing between LVCs (7) and phraseological idioms (8).

(7) *donijeti (dònijeti)* 'bring' ≈ *odlučiti* 'decide' (idiom)
frame:

[2] Some authors consider light verbs semantically empty, insignificant, or vague (e.g., Jespersen, 1942; Poutsma, 1926; Grimshaw and Mester, 1998; Cattell, 1984), while others argue that they affect a sentence's meaning since both verb and noun choices are constrained (Wierzbicka, 1982; Butt, 1995). We agree with the later perspective; therefore, we provide sense definitions even for semantically bleached verbs.

[3] PropBank (Hwang et al., 2010) uses the label ARGM-LVB for a light verb and ARG-PRX (ARGument-Predicating eXpression) for a predicative noun. They treat a light verb and true predicate, as they refer to the predicative noun, as a single predicating unit (REL), which assigns semantic roles by combining the arguments from both the light verb and the noun.

example: *donijeti odluku* ‘bring a decision = make a decision’

(8) *donijeti (dònijeti)* ‘bring’ ≈ *dati kome tko nije uložio nikakav trud* ‘to give to someone who has not made any effort’ (idiom)
frame:
example: *donijeti na tanjuru* ‘to bring on a plate = to hand (something) to someone on a silver platter’

In e-Glava (Birtić, Brač, and Runjaić, 2017), only psych-verbs are available; therefore, it does not currently contain LVCs.

Regarding general language online dictionaries, LVCs are listed as a separate sense of the verbs only in *Mrežnik* (Hudeček and Mihaljević, 2020). This 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 examples from corpora. The example of the description of the verb *napraviti* ‘make, do’ from *Mrežnik* is given in (9).

(9) Napraviti *se kao nepunoznačni glagol pojavljuje uz imenice i najčešće se može zamijeniti punoznačnim glagolom izrazno povezanim s odgovarajućom imenicom.*
‘*To make* as a light verb appears with nouns and can most often be replaced with a full verb that is connected to the corresponding noun.’
napraviti analizu
do/make analysis.ACC.SG
‘make an analysis’
Napraviti analizu znači analizirati što, pomno što proučiti.
‘Do/make (conduct) an analysis means to analyze something, to investigate something thoroughly.’
- *Također sam navela da prije izrade zakona treba napraviti detaljnu analizu i procjenu učinaka propisa.*
‘I also stated that before drafting a law, a detailed analysis and assessment of the regulation’s effects should be conducted.’

However, this is not systematically processed (see, e.g., *donijeti* ‘bring’).

In Verbion, the light verb is categorized as a distinct sense, based on conclusions that light verbs have semantic content beyond a mere functional role (e.g., Butt, 2010; Brugman, 2001; Jackendoff, 2007). It is paired with a predicative noun that gives it full meaning.[4] Verbs with a general meaning or highly schematic verbs, such as the verb *vršiti* ‘do, conduct’, are described with a generic definition “light verb that with a deverbal noun means to perform an activity” (12).

(12) **VRŠITI**
Eng. *do, conduct*
1. *nepunoznačni glagol koji s odglagolskom imenicom znači izvoditi ili obavljati kakvu radnju; obavljati*
‘light verb that with a deverbal noun means to perform or carry out some activity; carry out’
Semantic class: Verbs of General Activity
Frame: Intentionally_act
Example: *Inženjeri* ***su vršili*** *ispitivanja podmorja.*
‘The engineers conducted tests of the seabed.’

EX:	Inženjeri	**su vršili**	ispitivanja	podmorja
SYN:	NP	LV	NP	NP
MORPH:	NOM		ACC	GEN
SemR:	Agent		0	Theme
FE:	Agent			

LU:			analizu;
			čišćenje;
			dostavu,
			…	

The light verb is annotated with LV to differentiate it from semantically full verbs, which are marked with V. The question is whether a predicative noun bears a semantic role. One possible answer is that direct object NPs in LVCs do not bear a semantic role since the light verb is incapable of assigning one. However, due to the argument transfer, the direct object NPs transfer their argument structure to the argument structure of the light verb, resulting in the semantic role of the Theme being assigned to the object in the genitive case (Grimshaw and Mester, 1988;

[4] Deverbal nouns are listed using Word Sketch from (mainly) the Croatian Web Corpus (Ljubešić and Klubička, 2014).

Karimi-Doostan, 2004). Another answer could be that due to the argument sharing, both the light verb and the noun assign semantic roles (e.g., Culicover and Jackendoff, 2005; Butt 1995, 2010). At the moment, we have decided to label the semantic role of a predicative noun with 0. This choice also prompts the question of whether the NP in the genitive case functions as a noun complement or a verb complement (10). In this case, we treat it as a verb complement as when the LVC is substituted with the full verb *testirati* 'test', the NP in the genitive case is in the object positions in the accusative case (11).[5]

(10) *Inženjeri su vršili testiranje podmorja*.
engineers AUX did testing.ACC seabed.GEN
'The engineers conducted testing of the seabed.'
(11) *Inženjeri su testirali podmorje.*
engineers AUX test seabed.ACC
'The engineers tested the seabed.'

If a light verb accompanied by different nouns has a different sense, each sense will be recorded (e.g., *donijeti velike brige* 'bring big worries = cause great worry' – '*izazvati kakvu psihičku promjenu; zabrinuti* = cause a psychological change, to worry'; *donijeti zaradu* 'bring a profit' – '*doprinijeti čemu, biti koristan, često u materijalnom smislu* = to contribute to something, to be useful, often in a material sense').

4 Verbal idioms

In the general language dictionaries we analyzed, verbal idioms are a separate category within a lexicographic entry. In *Mrežnik*, an explanation of the verbal idiom and an example from the corpora are provided (12).

(12) *frazem: prodavati zjake*
idiom: 'twiddle one's thumbs
razg. Prodavati zjake znači dangubiti, besposličariti.
'colloq. *Prodavati zjake* means to laze around, to waste time.'
- Božo i Špela, naime, žive zajedno, no dok Špela radi i financira ih oboje, Božo po cijele dane prodaje zjake.
'Božo and Špela live together, but while Špela works and financially supports both of them, Božo spends his days twiddling his thumbs.'

Some online and printed valency dictionaries treat idioms as one of the verb's senses, while others record them separately from the verb's sense. VALLEX and CROVALLEX, which follows the methodology of VALLEX 1.0, record idioms as one of the senses of the main lemma. For example, in CROVALLEX, the verb *bacati* 'throw' includes as its sixth sense the idiom *bacati drvlje i kamenje (na nekoga)* (lit. throw wood and stones (at somebody) = scold somebody very much). The idiom's meaning is defined next to the sublemma with the label *idiom* in parentheses. The full idiom is quoted under the example section, but without the actual example, see (8) and (13).

(13) 6. *bacati (bàcati) ≈ jako grditi (idiom)*
'throw ≈ to scold severely'
-frame:
-example: *bacati drvlje i kamenje (na koga)*
'lit. throw wood and stones (at somebody) = to scold somebody severely'

In VALLEX, idioms are also recorded as one of the verb's senses, but with a detailed analysis of the idiom's components (similar to the semantic-syntactic characterization of other valency frames). For example, the ninth sense of the verb *házet*^{impf}, *hodit*^{pf} 'throw' includes a description of the idiom *hoditi se do gala* (lit. throw (one)self in / at a gala = to dress up) (14).[6]

(14) frame ACT^{obl}_{1} PAT^{obl}_{4} $\mathrm{DPHR}^{obl}_{do\ gala}$
example: hodit se do gala

The variable elements of the idiom are described as other valency elements by virtue of semantic roles (ACT, PAT), the case form expressed by a number and label indicating obligatoriness. The idiom's fixed element is invariably defined with an abbreviation DPHR (Dependent Phraseme),

[5] There is still no agreement among researchers on the project regarding the status of the NP in the genitive case (*podmorje* 'seabed'), as we are aware it can be treated as a noun complement (*testiranje* 'testing') due to the case assignment to *podmorje* 'seabed'.

[6] Special attention to idioms is given in the Polish Walenty (see Przepiórkowski et al., 2014).

followed by the exact form of the fixed element in subscript. In this case, it is the prepositional phrase *do gala* 'to the gala'.

On the other hand, e-Glava does not include idioms among the verb's senses but treats them in a separate section, after listing the verb's senses. There is a special section called *Čvrste sveze* 'Fixed phrases' under which all collocations and idioms connected with a certain verb are listed. This section contains the verbal idiom, its explanation, and an example from the corpora, but no detailed description of the syntax of the idiom, following VALBU (Schumacher et al., 2004) (15).

(15) *plašiti se svoje/vlastite sjene - biti vrlo plašljiv, pretjerano oprezan, biti kukavica*
'lit. to be afraid of one's/own shadow = to be very timid, overly cautious, to be a coward'
◊ *Za razliku od mnogih, koji se plaše vlastite sjene, Koki otvoreno progovara o svom poslu.*
'Unlike many who are cowards, Koki openly speaks about his work.'

In Verbion, idioms are recorded in a separate tab, as in e-Glava, but the components of the idioms are accompanied by a more detailed description, as demonstrated in (16). Although these descriptions are less detailed than those of the verb senses, as verbal idioms are not the primary focus of this phase of the project, they still provide valuable insights. Notably, the inclusion of translations of idioms into English is a significant contribution, especially given the current lack of online resources offering Croatian idioms with English translations, as far as we know.

(16) *donijeti na tanjuru = dati što komu tko nije uložio nikakav trud*
'bring on a silver platter = to give something to someone who hasn't put in any effort'
Example: *On joj je sve u životu* ***donio na tanjuru***.
'He brought her everything in life on a silver platter.'
Agent_NP_nom Theme_NP_acc Recipient_NP_dat
donijeti na tanjuru_VID

The verb, which can be conjugated and can appear in different tenses and moods, and the fixed part of the idiom are marked with the label VID (verbal idiom).

5 Conclusion

Investigating the processing of VMWEs in Croatian general language dictionaries and valency lexicons has highlighted the complexities of their recognition, classification, and description. The inconsistency in their treatment underscores the need for a more harmonized approach to the documentation and analysis of VMWEs. As showed in Section 2, the treatment of reflexive verbs in Croatian general language dictionaries is not unified either within a single dictionary or across dictionaries. The Czech VALLEX offers a compelling solution by treating all reflexive verbs as separate lemmas since they are distinct lexemes. By introducing different labels, we differentiate between reflexiva tantum (REFL tantum), derived reflexive verbs (REFL derived), and reciprocal reflexive verbs (REFL recipr). Currently, we lack resources for advanced classification and for the introduction of syntactic operations as in VALLEX, but this is planned for the future.

Light verb constructions present their own set of challenges, particularly in terms of, firstly, criteria for their recognition (which is not the topic of this paper), and for their semantic and syntactic representation. To distinguish between a semantically full verb and a light verb, we introduced the label V for a semantically full verb and LV for a light verb. By listing the most frequent lexemes in the position of a predicative noun and linking it to a full verb entry, we aim to improve the lexicon's usefulness.

Verbal idioms are included as a separate section, each accompanied by an explanation, a translation into English, an example from the corpus, and a syntactic and semantic description of its participants.

Other linguistic resources and the processing of VMEWs in Croatian can benefit from a verb lexicon that contains clearly marked and described reflexive verbs, light verb constructions, and, to some extent, verbal idioms.

Acknowledgments

This work has been supported by the Croatian Science Foundation under the project *Semantic-Syntactic Classification of Croatian Verbs* (SEMTACTIC) (IP-2022-10-8074) and by the European Union – NextGeneration EU. However, the views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Margarita Alonso Ramos. 2007. Towards the Synthesis of Support Verb Constructions: Distribution of Syntactic Actants between the Verb and the Noun. In Leo Wanner, editor, *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*. John Benjamins Publishing Company. 97–137.

Stjepan Babić, Dalibor Brozović, Ivo Škarić, and Stjepko Težak. 2007. *Glasovi i oblici hrvatskoga književnoga jezika*. Zagreb: Nakladni zavod Globus.

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The Structure of the FrameNet Database. *International Journal of Lexicography*, 16(3): 281–296.

Eugenia Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, and Marija Znika. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Branimir Belaj. 2001. Prototipno-kontekstualna analiza povratnih glagola u hrvatskom jeziku. *Suvremena lingvistika*, 51-52(1-2): 1–11.

Branimir Belaj. 2024. Croatian middle *se*-constructions. In M. Batinić Angster and M. Angster, editors, *The verbal kaleidoscope. Perspectives on the syntax and semantics of verbs*. University of Zadar, Zadar. 93–132.

Matea Birtić et al. 2012. *Hrvatski školski rječnik*. Zagreb: Školska knjiga.

Matea Birtić, Ivana Brač, and Siniša Runjaić. 2017. The main features of the e-Glava online valency dictionary. *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 43–62, Brno. Lexical Computing CZ s.r.o.

Claudia Brugman. 2001. Light verbs and polysemy. *Language Sciences*, 23: 551–578.

Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications, Stanford.

Miriam Butt. 2010. The Light Verb Jungle: Still Hacking Away. In M. Amberber, M. Harvey and B. Baker, editors, *Complex predicates in cross-linguistic perspective*. Cambridge University Press, Cambridge. 48–78.

Ray Cattell. 1984. *Composite Predicates in English*. Brill.

Silvie Cinková and Veronika Kolářová. 2006. Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*. L'udovít Štúr Institute of Li Linguistics of the SAS, Slovak National Corpus, Bratislava. 113–139.

Matthieu Constant 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, Berlin. ACL.

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

Charles J. Fillmore and Collin F. Baker. 2001. Frame Semantics for Text Understanding. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Workshop*, pages 1–6, Pittsburgh. Association for computational linguistics.

Jane Grimshaw and Armin Mester. 1988. Light Verbs and Θ-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.

Jena D. Hwang, Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. PropBank Annotation of Multilingual Light Verb Constructions. In *Proceedings of the Fourth Linguistickett Annotation Workshop, ACL 2010*, pages 82–90, Uppsala. Association for Computational Linguistics.

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

Otto Jespersen. 1942. *A Modern English Grammar on Historical Principles. Part VI. Morphology*. Ejnar Muksgaard, Copenhagen.

Gholamhossein Karimi-Doostan. 2005. Light verbs and structural case. *Lingua*, 115: 1737–1756.

Václava Kettnerová. 2023. Valency structure of complex predicates with Light Verbs. The case of Czech. In Anna Pompei, Lunella Mereu, and Valentina Piunno, editors, *Light Verb Constructions as Complex Verbs*. De Gruyter Mouton, Berlin. 19–43.

Václava Kettnerová, Markéta Lopatková, and Anna Vernerová. 2022. Reflexives as Part of Verb

Lexemes in the VALLEX Lexicon. *The Prague Bulletin of Mathematical Linguistics*, 119: 37–66.

Václava Kettnerová, Markéta Lopatková, Eduard Bejček, and Petra Barančíková. 2018. Enriching VALLEX with Light Verbs: From Theory to Data and Back Again. *The Prague Bulletin of Mathematical Linguistics*, 111: 29–56.

Václava Kettnerová, Petra Barančíková, and Marketa Lopatková. 2016. Lexicographic Description of Czech Complex Predicates: Between Lexicon and Grammar. In *Proceedings of the XVII EURALEX International Congress*, pages 881–892, Tbilisi. Ivane Javakhishvili Tbilisi State University.

Václava Kettnerová and Marketa Lopatková. 2013. The Representation of Czech Light Verb Constructions in a Valency Lexicon. In *Proceedings of the Second Conference on Dependency Linguistics, Depling 2013*, pages 147–156, Prague. Charles University in Prague.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, and Svetlozara Leseva. 2016. Semi-automatic Compilation of the Dictionary of Bulgarian Multiword Expressions. In *Proceedings of the GLOBALEX 2016: Lexicographic Resources for Human Language Technology*, pages 86–95, Portorož.

Beth Levin. 1993. *English Verb Classes and Alternations*. Chicago – London: The University of Chicago Press.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC - Web Corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg. Association for Computational Linguistics.

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

Ivana Oraić Rabušić. 2018. *Struktura povratnih glagola i konstrukcije sa* se *u hrvatskom jeziku*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Petya Osenova and Kiril Simov. 2018. The data-driven Bulgarian Wordnet: BTBWN. *Cognitive studies / Études cognitives*, 18: 1–18.

Hendrik Poutsma. 1926. *A Grammar of Late Modern English. For the use of continenta, especially Dutch, students. Part II. The Parts of Speech*. Noordhoff, Groningen.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2785–2792, Reykjavik. European Language Resources Association.

Dragutin Raguž. 2010. *Gramatika hrvatskoga jezika*. Zagreb: Vlastito izdanje.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing 2002, Lecture Notes in Computer Science, Volume 2276*, pages 1–15, Mexico City. Springer.

Agata Savary et al. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 31–47, Valencia. Association for Computational Linguistics.

Helmut Schumacher, Jacqueline Kubczak, Renate Schmidt, and Vera de Ruiter. 2004. *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Narr Verlag Tübingen.

Josip Silić and Ivo Pranjković. 2005. *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.

Anna Wierzbicka. 1982. Why Can You Have a Drink When You Can't *Have an Eat?. *Language*, 58(4): 753–799.

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