Assigning FrameNet Frames to a Croatian Verb Lexicon

Ivana Brač and **Ana Ostroški Anić**
Institute for the Croatian Language
ibrac@ihjj.hr; aostrosk@ihjj.hr

Abstract

This paper presents the Croatian verb lexicon Verbion that describes verbs on multiple levels. The semantic level includes verb senses, corresponding semantic classes according to VerbNet and WordNet, as well as semantic frames based on FrameNet. Each verb sense is linked to one or more valency frames, which include corpus-based examples accompanied by syntactic, morphological, and semantic analyses of each argument. This study focuses on assigning FrameNet frames to the verb *misliti* 'think' and its prefixed forms. Based on 170 manually annotated sentences, the paper discusses the advantages and challenges of assigning semantic frames to Croatian verbs.

1 Introduction

Verbs have been extensively analyzed in various linguistic resources as they are traditionally regarded to be the core element of a sentence. Different resources examine different aspects of verbs, focusing on semantics, e.g., WordNet (Fellbaum, 1998), FrameNet (Ruppenhofer et al., 2016); both semantics and syntax, e.g., VerbNet (Kipper, Dang, & Palmer 2000), PropBank (Bonial et al., 2010); or semantics, syntax, and morphology, e.g., VALLEX (Lopatková et al., 2021), Walenty (Przepiórkowski et al., 2014), CROVALLEX (Mikelić Preradović, 2020), e-Glava (Birtić, Brač, & Runjaić, 2017), CroaTPAS (Marini & Ježek, 2021). Despite being developed within different theoretical frameworks, these resources could have benefitted significantly from cross-mapping or linking. For instance, VALLEX tried to enhance its description by introducing information from FrameNet (Kettnerová, Lopatková, & Bejček, 2012), while the Unified Verb Index integrated links from diverse NLP projects such as VerbNet, PropBank, FrameNet, OntoNotes (Hovy et al. 2006), and the SynSemClass Lexicon (Straková et al., 2023).

Due to the lack of such resources for Croatian (except WordNet to a certain extent, Šojat, 2012), automatic linking is not currently feasible. However, a database is being developed to integrate various approaches and data into a comprehensive verb description. In this paper, we introduce a Croatian verb lexicon that describes verbs on several levels (Section 2) and, using the verb *misliti* 'think' and its prefixed forms, i.e. *pomisliti* 'think, have a thought', razmišljati$_{IMPF}$/razmisliti$_{PERF}$ 'think, think over, ponder', *smisliti* 'think of, come up with', *zamisliti* 'imagine, evision', *promisliti* 'think through, reflect on', and *izmisliti* 'make up, invent, fabricate', we reflect on the advantages and challenges of applying Frame Semantics to the description of verbs in Croatian (Sections 3 and 4).

The paper addresses the following key research questions: 1. What are semantic similarities and differences between the Croatian verb *misliti* 'think' and its prefixed forms? 2. Are semantic frames from the Berkeley FrameNet applicable to a description of Croatian verbs of thinking? As the result of the analysis and annotation of 170 sentences, new semantic frames are introduced in the Croatian data, and new lexical units suggested to be added to the existing frames.

2 Verb Lexicon Verbion

Verbion is a Croatian verb lexicon that will be publicly available by the end of 2027 through an online search interface offering advanced search options across various linguistic categories. XML data will be made available to researchers upon request for scientific purposes. In the first phase of the project, the 500 most frequent verbs will be described on several levels. On the first level, for each verb, its morphological aspect, an aspectual pair, a morphological block containing different

tenses and moods, its English equivalent, idioms, and senses are determined.

On the second level, each sense is associated with the VerbNet's (Kipper-Schuler 2005; Kipper et al., 2008) and WordNet's (Fellbaum, 1998) semantic classes. As is well known, the starting point for VerbNet's semantic classes is Levin's classification (1993), which is based on syntactic alternations, assuming that a verb's syntactic behavior reflects its semantics. However, some classes and subclasses are missing from Levin's classification (1993) since she focused on verbs with noun and prepositional phrase arguments. Consequently, VerbNet introduced more than 80 classes and subclasses (Dorr 1997; Korhonen & Briscoe 2004; Kipper et al. 2008) to account for this gap. Problems with Levin's classification arose even in the case of verbs with relatively straightforward sense description, like the verb *think*. In Levin's classification, *think* belongs to the class of verbs with predicative complements, specifically, to the subclass of *declare* verbs. In contrast, VerbNet classifies it into three different classes (*consider-29.9-2*, *focus-87.1-1*, and *wish-62*). In Verbion, hierarchically organized semantic classes are introduced, preserving Levin's original classes while incorporating VerbNet's subclasses and newly established classes. WordNet's classification, on the other hand, is based solely on semantic criteria and contains fewer classes, i.e., stative verbs and 14 action verb classes.

Different verb senses can belong to different semantic classes. For example, two senses of the verb *misliti* 'think' – 'to have someone or something in mind' and 'to have an opinion about someone or something' – belong to the *focus-87.1-1* subclass, while in the sense 'to take care of someone or something, carry, worry', it falls under the *caring-75.2* subclass and WordNet's *verb.emotion* class. On the other hand, in the sense 'to intend to do something', it belongs to the *intend-61.2* subclass and WordNet's *verb.cognition*. The second level of verb description also contains definitions in Croatian and English, Croatian synonyms and English equivalents of the defined verb sense, and a semantic frame. For each verb sense, the corresponding FrameNet's frame is identified, and for each participant, the appropriate frame element is determined. Frames in Verbion are linked to Berkeley's FrameNet and Croatian FrameNet, which is being developed.

On the third level of description, each sense is associated with one or more valency frames, which include examples from corpora, their translation into English, and an analysis of participants at three levels: syntactic, morphological and semantic levels. At the syntactic level, each participant is marked with syntactic phrase type, similar to VerbNet, but with a few modifications (e.g., CP instead of S). Since Croatian cases are morphologically realized, the morphological realization of syntactic phrase is specified. For the semantic description of the participants, slightly modified semantic roles from VerbNet are used.

This approach aims to make the description of verbs as comprehensive as possible, and one of the means is incorporating the frame-semantic framework used to define verbs following the principles of Frame Semantics (Fillmore, 1985; Ruppenhofer et al., 2016). There have been many extensions of FrameNet to other languages, many of which have been created by expanding the original FrameNet with translations into their language, e.g., the Spanish FrameNet (Subirats, 2009). Others resorted to merging the FrameNet model with the existing resources, e.g. the Czech FrameNet, developed by linking Verbalex to FrameNet (Materna & Pala, 2010). Of Slavic languages, Bulgarian FrameNet has been by far the most developed (Koeva, 2010).

3 Methodology

To determine verb senses, Croatian online dictionaries (`https://hjp.znanje.hr/`; `https://rjecnik.hr/`) were consulted, as well as web corpora since some senses may be missing from the dictionaries. Data for the analysis was extracted from two Croatian general language web corpora, hrWaC (Ljubešić & Klubička, 2014) and CLASSLA (Ljubešić & Kuzman, 2024), based on manual analysis of random sample of 300 sentences for each analyzed verb. First, concordances had been analyzed in Sketch Engine (Kilgarriff et al., 2014) to identify common valency frames for each verb. Word Sketches were then used to check any potentially missing valency frames in the random sample, as well as regular expressions for more targeted searching.

In the second phase, ten sentences per each verb's sense were manually selected and annotated for FrameNet's semantic frames, applying the Berkeley FrameNet 1.7, which yielded 170 sentences. Although annotation was done by two

annotators, inter-annotator agreement was not measured at this stage as the focus of the task was to perform qualitative analysis and create guidelines for future annotation work.

4 FrameNet and Frames of Thinking

The verb *misliti* 'think', as the central member of the category of verbs of thinking, can be used in Croatian to express at least four senses: 1. 'to form or have someone or something in mind,' 2. 'to have an opinion about someone or something,' 3. 'to take care of someone or something,' and 4. 'to intend to do something.' The annotation of sentences extracted from corpora showed that these senses can be linked to four semantic frames, i.e. `Awareness`, `Cogitation`, `Opinion`, and `Regard`, but the comparison of Croatian senses of the verb *think* to the different senses of the lexical unit (LU) *think* in the Berkeley FrameNet shows certain differences in the conceptualization. In Croatian, the most frequent sense of the verb, 'to have an opinion about someone or something,' covers two senses of the LU *think* in FrameNet: one realized in the frame `Opinion`, and the other in the frame `Regard`.

(1) *Mislim* da je [strah od smrti TOPIC] [prirodan OPINION]. CNI COGNIZER
'[I COGNIZER] *think* the [fear of death TOPIC] is [natural OPINION].'
(2) *Mislite* [o meni EVALUEE] [što god hoćete JUDGEMENT]. CNI COGNIZER
'*Think of* [me EVALUEE] [whatever you want JUDGEMENT].'CNI COGNIZER

In (1), *think* evokes the frame `Opinion` as the COGNIZER (expressed as the 1st person singular form of the verb) holds an OPINION of a certain TOPIC, whereas in (2), the COGNIZER (expressed as the 2nd person plural imperative form of the verb) should be annotated as the frame element (FE) of `Regard` because the COGNIZER has a JUDGEMENT of an EVALUEE. Since corpus examples showed that there was no difference in valency patterns in Croatian between the two uses of this sense – holding an opinion about something or someone and having a judgement – both instances are defined in the Verbion database as belonging to the frame `Opinion`.

When used in its third sense, 'to take care of someone or something,' the verb *misliti* 'think' evokes the frame of having concern for someone, as in (3):
(3) Nismo sebični, *mislimo* o svim žrtvama rata, ne gledajući na vjeru, naciju i uniformu.
'We are not selfish; we *think* of all the victims of war, regardless of faith, nationality, or uniform.'

Although this sense of *think* is not described in FrameNet, and there is no corresponding frame defined which could encompass it, the sense is nevertheless attested in English, as evidenced in this example given in Merriam-Webster: *I must think first of my family*. It is therefore justified to introduce a new frame `Take_care_of`, that also includes other lexical units, e.g., *care* (n.), *care for* (v.), *take care* (v.), *concern* (n.), etc. Finally, using *misliti* 'think' in the sense of 'having a plan or intention to do something' is the second most frequent use of the verb *think* (4):

(4) Ako *misliš* [oženiti se GOAL], napravi to dok si mlad jer kasnije nećeš htjeti. CNI AGENT
'If [you AGENT] *think of* [getting married GOAL], do it while you're young because later you won't want to.

Examples with *misliti* used in this sense are annotated in the frame `Purpose`, which underlines the role of the AGENT, although a frame for expressing intent would have been better suited for the meaning. FrameNet does not list the sense 'plan to do something' for the LU *think*, although it is confirmed in usage, as in *I've been thinking of buying a boat*.

Prefixed verbs related to the verb *misliti* 'think' align with different frames, showing how prefixes encode subtle semantic distinctions. Slavic prefixes modify both the aspect and the semantic focus of the verb. In contrast, English tends to use separate verbs or verb phrases to convey similar nuances (cf. Svenonious, 2005). Slavic prefixes are not empty prefixes (Janda, 1986; Belaj, 2008); therefore, they serve as meaning modifiers rather than mere aspect markers, which can be shown in the next examples.

The first, canonical sense of the verb *misliti* 'think', 'to form or have someone or something in mind' (as in *Mislio sam o tebi.* 'I've been thinking about you,') in Croatian is commonly expressed with the perfective verb *razmišljati* 'think, think about, think through, think over, ponder.' Both senses evoke the frame `Cogitation`, in which the COGNIZER thinks about a TOPIC over a period of

time. This verb emphasizes duration, intensity and excessiveness of the process of thinking. Duration is all the more underlined by the use of the perfective verb like *razmišljati*. The prefix *raz-* typically signifies a transition of the trajector's state from compact to a dispersed one (Belaj, 2004, 2008). In the context of thinking, this means that thoughts are initially directed towards the object as a whole, and then different aspect or every part of it are thought through. The trajector is broken into smaller parts and analyzed from different angles.

`Cogitation`, was also used to annotate the first sense of another prefixed verb, *pomisliti* 'think, think about, have a thought.' *Pomisliti* can either stand for 1. 'to momentarily form a thought or create an idea that often arises as an initial reaction or intuitive impression about something,' and 2. 'to recall someone or something.' Sentences expressing the second sense are annotated using the FEs of the frame `Remembering_experience`, but there is no appropriate frame in FrameNet for the sense of momentarily forming a thought or creating an idea, as in example (5) and (6):

(5) Za scenarij je odmah *pomislila* da je briljantan.
'She immediately *thought* the script was brilliant.'
(6) Ni u kojem trenutku nemojte *pomisliti* na šminkanje prije odlaska na plažu.
'At no point should you *think about* putting on makeup before going to the beach.'

When used to form verbs, the prefix *po-* can stand (among its other uses) for the beginning of the activity expressed by the verb, as well as to express that the activity is completely finished. In the verb *pomisliti*, it highlights the moment in which the thought is created. These subtle differences between the Croatian verb *pomisliti* and its English equivalent *think* can be seen in (7), where the implied meaning of the Croatian sentence is 'I have never even had one bad thought about my mother,' which is not present in the English translation.

(7) Nikada nisam ništa loše *pomislio* o mojoj mami.
'I have never *thought* anything bad about my mom.'

Examples like (5), (6) and (7) have been annotated using the `Cogitation` frame as it is the closest frame containing the most relevant frame elements. The aspect of a "sudden" thinking in the process, or the moment that the thinking starts is annotated using the FE MANNER, e.g. *immediately* in (5), *at no point* in (6) and *never* in (7) are all annotated as FE MANNER in the frame `Cogitation`.

With the verbs *smisliti* 'think of, come up with' and *izmisliti* 'to make up, invent, fabricate,' the process of thinking leads to the creation of an idea. The prefix *iz-* denotes extraction or emergence, much like *s-*, but with a key difference: *iz-* typically implies that the landmark is a container, whereas *s-* suggest a surface. This distinction can be conceptualized as ideas coming off the top of one's head versus being deeply extracted from the mind (cf. Krawczak & Kokorniak, 2012, p. 451). With the verb *smisliti*, thinking is solution-oriented, focusing on devising a concreate idea or plan. Meanwhile, *izmisliti* implies the act of bringing an idea into existence, whether real or fictional. However, in both cases, the result of the process of thinking emerges from one's mind (cf. Dickey, 2005, p. 37). However, the verb *smisliti* belongs to the semantic frame `Coming_up_with` (8), which highlights the mental effort involved in generating a solution or plan, while *izmisliti* fits into the frame `Achieving_first` (9), which highlights the creation of something novel or original, often with an element of innovation.

(8) [Ime IDEA] je *smislio* [njezin brat COGNIZER].
'[Her brother COGNIZER] *came up* [with the name IDEA].'
(9) [Europljani COGNIZER] su *izmislili* [kotač NEW_IDEA].
'[Europeans COGNIZER] *invented* [the wheel NEW_IDEA].'

The prefix *za-* has inchoative meaning and in the case of the verb *zamisliti* 'imagine, envision', it expresses the beginning or the setting up of an idea, which often involves creativity or visualization.

(10) *Zamislite* [savršeno mjesto za odmor CONTENT].
CNI COGNIZER
'*Imagine* [a perfect place to relax CONTENT].' CNI COGNIZER

Example (10) is therefore annotated using the `Awareness` frame, in which the idea or visualization that the COGNIZER has serves as the CONTENT of the act of cognition.

5 Conclusion

In many less- and under-resourced languages, the challenges of developing complex lexical resources are all the greater as there is a lack of more fundamental linguistic resources (e.g., learners' monolingual dictionaries, monitor corpora or a thesaurus, to name a few), that will probably never be created.

The Verbion database aims to fill that void in Croatian by merging several linguistic approaches in order to provide an all-encompassing description of most frequent verbs in Croatian. Apart from focusing on the presentation of their arguments structure, Verbion also includes a semantic description of verbs classified into semantic classes. The analysis of verbs of thinking presented here proves that different lexical resources can be successfully merged with minimal adjustments. 170 sentences containing 8 verbs of thinking in Croatian were annotated using 8 semantic frames from the original FrameNet data, and compared to their English translations to establish links with equivalent frame elements. In most examples, existing FEs were the exact match to annotate Croatian lexical units, or could have been well used to account for a very similar meaning. One new semantic frame needed to be defined, `Take_care_of`, which did not exist in FrameNet to describe situations when an AGENT *looks after* someone, *takes care of* someone, or *thinks of* someone in the same context. In certain examples, a decision had to be made whether to go for a more granular or schematic description of the verb's sense, e.g. for the senses of opinion and judgement of the verb *misliti* 'think'. A finer semantic description will be kept in future Croatian FrameNet, as opposed to Verbion that does not exclusively rely on semantic frames for verb description.

This analysis will serve as the model for developing benchmarks for the validation of automatic frame assignment, which is particularly important for languages like Croatian, with rich morphology. Scarce online resources, particularly semantically based lexical resources, present an obstacle in the development of LLM-based applications for Croatian and other less-resourced languages. The creation of verified and valid frame-based lexical resources will certainly improve the efficiency of the existing LLMs, and help in their applications.

Acknowledgments

This work was created as part of the projects *Semantic-syntactic classification of Croatian verbs* – SEMTACTIC (IP-2022-10-8074), funded by the Croatian Science Foundation, and *Semantic Frames in the Croatian Language* and *Croatian Verb Valencies* funded by the European Union – NextGenerationEU.

References

Branimir Belaj. 2008. *Jezik, prostor i konceptualizacija. Shematična značenja hrvatskih glagolskih prefiksa*. Sveučilište Josipa Jurja Strossmayera u Osijeku, Osijeka.

Branimir Belaj. 2004. Značenjska analiza hrvatskoga glagolskog prefiksa *raz-* i njegovih alomorfa *ras-*, *raš-*, *raž-*, *raza-*, *ra-*. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 30:1–16.

Matea Birtić, Ivana Brač, and Siniša Runjaić. 2017. The main features of the e-Glava online valency dictionary. In *Electronic Lexicography in 21st Century Proceedings of eLex 2017 Conference*. Lexical Computing, Brno, pages 43–62.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. *PropBank Annotation Guidelines*. Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder.

Stephen M. Dickey. 2005. *S-/Z-* and the Grammaticalization of Aspect in Slavic. *Slovenski jezik – Slovene Linguistic Studies*, 5:1–55.

Bonnie J. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–325.

Christiane Fellbaum. 1998. A Semantic Network of English Verbs. In *WordNet: an electronic lexical database*. The MIT Press, Cambridge – London, pages 69–104.

Charles J. Fillmore. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica* 6. 222−254.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*. Association for Computational Linguistics, pages 57–60.

Laura A. Janda. 1986. *A Semantic Analysis of the Russian Verbal Prefixes za-, pere-, do-, and ot-*. Slavistische Beitrage, Munchen.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. FrameNet II: Extended Theory and Practice, https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

Václava Kettnerová, Markéta Lopatková, and Eduard Bejček. 2012. Mapping Semantic Information from FrameNet onto VALLEX. *The Prague Bulletin of Mathematical Linguistics*, 97:23–41.

Adam Kilgarriff, Vít Baisa, Jan Busta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, P. Rychlý and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1:7–36. https://doi.org/10.1007/S40607-014-0009-9

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources & Evaluation*, 42:21–40.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. *AAAI-Proceedings*, pages 691–696.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Dept., University of Pennsylvania.

Svetla Koeva. 2010. *Bulgarian FrameNet*. Institute for Bulgarian Language Prof. Lyubomir Andreychin, Sofia.

Anna Korhonen, and Ted Briscoe. 2004. Extended Lexical-Semantic Classification of English Verbs. In *Proceedings of HLT/NAACL'04 Workshop on Computational Lexical Semantics*. Association for Computational Linguistics, Boston, pages 38–45.

Karolina Krawczak, and Iwona Kokorniak. 2012. A corpus-driven quantitative approach to the construal of Polish *think*. *Poznań Studies in Contemporary Linguistics*, 48(3):439–472.

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

Markéta Lopatková, Václava Kettnerová, Anna Vernerová, A., Eduard Bejček, and Zdenek Žabokrtský. 2021. *Valenční slovník českých sloves VALLEX*. UFAL Technical Report TR-2021-68.

Nikola Ljubešić, and Taja Kuzman. 2024. CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Gentre Annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, pages 3271–3282.

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)*. Association for Computational Linguistics, Gothenburg, pages 29–35.

Constanza Marini, Elisabetta Ježek. 2021. CROATPAS: A Lexicographic Resource for Croatian Verbs and its Potential for Croatian Language Teaching. In *Proceedings of the 19th EURALEX International Congress*. Democritus University of Thrace, Alexandroupolis, pages 529–534.

Jiří Materna and Karel Pala. 2010. Using ontologies for semi-automatic linking VerbaLex with FrameNet. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/867_Paper.pdf

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

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)*. ELRA, Reykjavik, pages 2785–2792.

Jana Straková, Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2023. Extending an Event-type Ontology: Adding Verbs and Classes Using Fine-tuned LLMs Suggestions. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 85–95.

Carlos Subirats. 2009. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In *Multilingual FrameNets in Computational Lexicography. Methods and Applications*, Hans C. Boas (ed.), pages 135–162. De Gruyter Mouton, Berlin – New York, doi:10.1515/9783110212976.2.135.

Peter Svenonius. 2005. Slavic prefixes inside and outside VP. *Nordlyd*, 32(2):205–253.

Krešimir Šojat. 2012. Struktura glagolskog dijela Hrvatskog WordNeta. *Filologija*, 59:153–172.