CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian

Costanza Marini
University of Pavia
Department of Humanities
costanza.marini01@universitadipavia.it

Elisabetta Ježek
University of Pavia
Department of Humanities
jezek@unipv.it

Abstract

The goal of this paper is to introduce CROATPAS, the Croatian sister project of the Italian Typed-Predicate Argument Structure resource (TPAS[1], Ježek et al. 2014). CROATPAS is a corpus-based digital collection of verb valency structures with the addition of semantic type specifications (SemTypes) to each argument slot, which is currently being developed at the University of Pavia. Salient verbal patterns are discovered following a lexicographical methodology called Corpus Pattern Analysis (CPA, Hanks 2004 & 2012; Hanks & Pustejovsky 2005; Hanks et al. 2015), whereas SemTypes – such as [HUMAN], [ENTITY] or [ANIMAL] – are taken from a shallow ontology shared by both TPAS and the Pattern Dictionary of English Verbs (PDEV[2], Hanks & Pustejovsky 2005; El Maarouf et al. 2014). The theoretical framework the resource relies on is Pustejovsky's Generative Lexicon theory (1995 & 1998; Pustejovsky & Ježek 2008), in light of which verbal polysemy and metonymic argument shifts can be traced back to compositional operations involving the variation of the SemTypes associated to the valency structure of each verb. The corpus used to identify verb patterns in CROATPAS is the Croatian Web as Corpus (hrWac 2.2, RELDI PoS-tagged) (Ljubešić & Erjavec 2011), which contains 1.2 billion types and is available on the Sketch Engine[3] (Kilgarriff et al. 2014). The potential uses and purposes of the resource range from multilingual pattern linking between compatible resources to computer-assisted language learning (CALL).

1 Introduction

Nowadays, we live in a time when digital tools and resources for language technology are constantly mushrooming all around the world. However, we should remind ourselves that some languages need our attention more than others if they are not to face – to put it in Rehm and Hegelesevere's words – "a steadily increasing and rather severe threat of digital extinction" (2018: 3282).

According to the findings of initiatives such as the META-NET White Paper Series (Tadić et al. 2012; Rehm et al. 2014), we can state that Croatian is unfortunately among the 21 out of 24 official languages of the European Union that are currently considered *under-resourced*. As a matter of fact, Croatian "tools and resources for […] deep parsing, machine translation, text semantics, discourse processing, language generation, dialogue management simply do not exist" (Tadić et al. 2012: 77). An observation that is only strengthened by the update study carried out by Rehm et al. (2014), which shows that, in comparison with other European languages, Croatian has *weak to no support* as far as text analytics technologies go and only *fragmentary support* when talking of resources such as corpora, lexical resources and grammars.

In this framework, a semantic resource such as CROATPAS could play its part not only in NLP, (e.g. multilingual pattern linking between other existing compatible resources), but also in automatic machine translation, computer-assisted

[1] http://tpas.fbk.eu (last visited on July 12th 2019)

[2] http://pdev.org.uk (last visited on July 12th 2019)

[3] https://www.sketchengine.eu/ (last visited on July 12th 2019)

language learning (CALL) and theoretical and applied cross-linguistic studies.

The paper is structured as follows: first a detailed overview of the resource is presented (Section 2), followed by its theoretical underpinnings (Section 3) and a summary of the Croatian-specific challenges we faced while building the resource editor (Section 4). An overview of the existing related works is given in Section 5. Finally, Section 6 hints at the creation of a multilingual resource linking CROATPAS, TPAS (Italian) and PDEV (English) patterns and explores CROATPAS's potential for computer-assisted L2 teaching and learning.

2 Resource overview

CROATPAS, i.e. the Croatian Typed-Predicate Argument Structure resource, is the Croatian equivalent of the Italian TPAS resource (Ježek et al. 2014) and is a corpus-derived collection of Croatian verb argument structures, whose argument slots have been annotated using semantic type specifications (SemTypes).

The first version of the resource is currently being developed at the University of Pavia with the technical assistance of *Lexical Computing Ltd*. in the person of Vìt Baisa and will be released in 2020 through an Open Access graphical user interface on the website of the Language Centre of the University of Pavia (CLA)[4].

CROATPAS contains a sample of 100 medium-frequency Croatian verbs, whose Italian translational counterparts are already available in the TPAS resource: 26 of these verbs are Croatian translational equivalents of Italian "coercive verbs", i.e. verbs that instantiate metonymic shifts in one of their senses (Ježek & Quochi 2010), while the remaining 74 are Croatian translational equivalents of a sample of Italian *fundamental verbs*, i.e. verbs belonging to that group of approximately 2000 lexemes deemed essential for communicating in Italian and that can be found in any sort of text (De Mauro 2016).

Our 74-verbs sample was selected as follows: we first extracted the frequency counts for all the 452 fundamental verbs on De Mauro's list from a reduced version of the ItWAC (Baroni & Kilgarriff, 2006), which contains over 900 million tokens and is available on the Sketch Engine (Kilgarriff et al. 2014). We then selected our 74 Italian candidates around the median frequency value after taking out the first and the last 20 verbs on the list. Finally, the Croatian translational equivalents for these verbs were chosen using the 2017 Zanichelli Italian/Croatian bilingual dictionary *Croato compatto*, edited by Aleksandra Špikić.

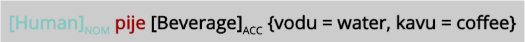
The theoretical framework the resource relies on is Pustejovsky's Generative Lexicon theory (1995 & 1998; Pustejovsky & Ježek 2008), in light of which verbal polysemy and metonymic shifts can be traced back to compositional operations involving the contextual variation of the SemTypes associated to the valency structure of each verb.

CROATPAS rests on four key-components, namely:

1) a representative corpus of Croatian;
2) a shallow ontology of SemTypes;
3) a methodology for corpus analysis;
4) adequate corpus tools.

As for the first component, the corpus used to identify verb patterns is the Croatian Web as Corpus (hrWac 2.2, RELDI PoS-tagged) (Ljubešić & Erjavec, 2011), containing 1.2 billion types and available on the Sketch Engine (Kilgarriff et al. 2014). We chose to work with the Croatian Web as Corpus since the reference corpus for the Italian TPAS resource is a reduced version of the Italian Web as Corpus (Baroni & Kilgarriff, 2006), so as to make the two resources as comparable as possible.

As for the shallow ontology of Semantic Type labels, CROATPAS is based on the same hierarchy shared by TPAS and the PDEV project of 180 SemTypes, which originates from the *Brandeis Shallow Ontology* (BSO) (Pustejovsky et al. 2004) and its initial 65 labels. As pointed out by Ježek (2014: 890), SemTypes "are not abstract categories but semantic classes discovered by generalizing over the statistically relevant list of collocates that fill each position". For example, the Croatian lexical set for the SemType [BEVERAGE] in the context of the verb pair PITI/POPITI (= TO DRINK, *imperfective/perfective*) contains, among others: {vodu = water, kavu = coffee, koktel = cocktail, vino = wine, čaj = tea, pivo = beer, limonadu = lemonade}, as shown in the following pattern string from the resource.



[Human]$_{NOM}$ pije [Beverage]$_{ACC}$ {vodu = water, kavu = coffee}

Figure 1 – One of the pattern strings of PITI

[4] https://cla.unipv.it/?page_id=53723 (last visited on July 12th 2019)

The corpus analysis methodology used for both TPAS and CROATPAS is a lexicographical methodology called Corpus Pattern Analysis (CPA, Hanks 2004 & 2012; Hanks & Pustejovsky 2005; Hanks et al. 2015), which is based on the Theory of Norms and Exploitations (TNE, Hanks 2004, 2013). TNE divides word uses in two main classes: conventional uses (*norms*) and deviations from the norms (*exploitations*). CPA's potential lies in that it does not try to identify meaning in isolation, but rather associates it with prototypical contexts, thus focusing on the norms. The standard CPA procedure requires:

1) sampling concordances for each verb
2) identifying its typical patterns – i.e. senses – while going through the corpus lines
3) assigning SemTypes to the argument slots in each pattern
4) assigning the sampled concordance lines to the identified patterns

This last operation is possible because both the TPAS and CROATPAS editors are linked to their respective language-specific corpora through the Sketch Engine (Kilgarriff et al. 2014), which proves once again to be the perfect tool for lexicographic work.

The resource will be evaluated through IAA on pattern identification for a sub-sample of the verb inventory, following the methodology proposed by Cinkova et al. (2012).

3 Generative Lexicon Theory

As pointed out by Hanks (2014: 1), the CPA methodology relies theoretically on the Theory of Norms and Exploitations (TNE), which has its roots in Sinclair's work, but is also influenced by Pustejovsky's Generative Lexicon Theory (1995 & 1998; Pustejovsky & Ježek 2008), thus bridging the gap between corpus linguistics and semantic theories of the lexicon.

In his theory, Pustejovsky tries to account for the semantic richness of natural language focusing on the compositional aspects of lexical semantics. According to this framework, lexical meaning is not an intrinsic feature of lexical items, but is generated by means of their contextual interaction, following the so-called *principles for strong compositionality*. As outlined in Ježek (2016: 78), these principles operate at a sub-lexical level targeting specific aspects of word meaning – such as SemTypes – and are able to provide different interpretations for a wide range of lexical phenomena.

The *principle of co-composition*, for instance, offers an alternative take on verbal polysemy with respect to traditional accounts. If we consider lexical items expressing verb arguments to be as semantically active and influential as the verb itself (Pustejovsky 2002: 421), we do not need to think of verbs as polysemous, but rather conceive their meaning as contextually defined by the SemTypes of the surrounding arguments. For instance, if we apply this reasoning to the Croatian verb pair PITI/POPITI (= TO DRINK, *imperfective/perfective*), we can notice how its meaning changes depending on what is said to be "drunk", namely a [BEVERAGE] (1), a [DRUG] (2) or a {GOAL} (3).

(1) [[HUMAN$_{\text{NOM}}$] **PIJE** [[BEVERAGE]$_{\text{ACC}}$]
Djeca *ne piju* *kavu.*
Children don't drink coffee.

(2) [[HUMAN$_{\text{NOM}}$] **PIJE** [[DRUG]$_{\text{ACC}}$]
Većina ljudi *pije* *antibiotike* *na svoju ruku.*
Most people take antibiotics on their own initiative.

(3) [[HUMAN_FOOTBALL PLAYER]$_{\text{NOM}}$] **POPIJE** {GOL}
Pavić *je popio* *gol.*
Pavić failed to score a goal.

As for metonymic phenomena, in this framework they take the name of *semantic type coercions* (Pustejovsky 2002: 425; Pustejovsky & Ježek 2008, Ježek & Quochi 2010). Unlike co-composition instances, coercions do not cause shifts in verb meaning, but rather operate semantic type adjustments to the verb's selectional requirements within a given pattern. For instance, when a verb such as POPITI combines with a Direct Object with the semantic type [CONTAINER] in a context where it should select [BEVERAGE], it is instantiating a metonymic shift which enables us to interpret the given [CONTAINER] as the [BEVERAGE] itself, like in example (4).

(4) [[HUMAN$_{\text{NOM}}$] **POPIJE** [[CONTAINER]$_{\text{ACC}}$]
Stipe *je popio* *čašu.*
Stipe drank a glass.

4 Croatian-specific challenges

Being a Slavic language, Croatian displays a certain number of language-specific features, which had to be taken into account when setting up the new editor for CROATPAS, such as its case system, the consequent absence of prepositions when case markings are providing information on clause roles and verbal aspect. We implemented an editor which is proving to be able to tackle those challenges.

For instance, the following example (5) taken from the verb POSLATI (= TO SEND, *perfective*) shows how the addition of case markings as bottom-right indexes has proven essential to make the resource user-friendly: had they not been there, the absence of the preposition "to" in Croatian would have made Theme and Recipient morphologically undistinguishable from one another.

(5) [[HUMAN]$_{\text{NOM}}$] **POŠALJE** [[ARTEFACT]$_{\text{ACC}}$] [[HUMAN]$_{\text{DAT}}$]
Marija je poslala pismo gradonačelniku.
Marija sent a letter **TO** the mayor.

For what concerns sentence structure, like the acronym suggests, the Croatian Typed Predicate Argument Structure resource leans on valency theory, where no distinction is made between subject and obligatory complements, since they are all considered essential verb *arguments* (Ježek 2016: 112). However, the editors of both TPAS and CROATPAS still rely on traditional clause-role labels for the underlying syntactic annotation, thus distinguishing subjects from objects and other obligatory complements.

Also traditional Croatian grammar distinguishes between clause roles, but the classification is heavily influenced by the Croatian case system and the use of prepositions. Croatian makes use of seven morphological cases – nominative, genitive, dative, accusative, vocative, locative and instrumental – which go by the name of *padeži* (Barić et al. 1997: 101)[5]. *Subjects* are usually expressed by the nominative case (6) (*ibidem*, 421), apart from some logical subjects appearing in the dative case (7).

(6) *Ivan-Ø je simpatičan-Ø*
Ivan-**NOM** is nice-NOM
'Ivan is nice'

(7) *Vrti mi se*
(It) spins I.**DAT** REFL
'I feel dizzy'

Direct objects (*ibidem*, 431) are expressed either by the accusative (8) or the genitive case (9), in case the context calls for a partitive genitive (*ibidem*, 435).

(8) *Irin-a čita knjig-u*
Irina-NOM reads book-**ACC**
'Irina reads a book'

(9) *Hoćeš li kruh-a?*
(you) need Q bread-**GEN**
'Do you want some bread?'

Indirect objects are expressed either by the genitive (10), dative (11) or instrumental case (12) (*ibidem*, 436).

(10) *Bojim se smrt-i*
(I) fear REFL death-**GEN**
'I am afraid of death'

(11) *Veselim se Božić-u*
(I) rejoice REFL Christmas-**DAT**
'I look forward to Christmas'

(12) *Revolver-om je lako rukovati*
Revolver-**INS** (it) is easy to handle
'It is easy to handle a revolver'

Another distinction made in traditional Croatian grammar is the one between *non-prepositional* and *prepositional objects* (*ibidem*, 443): subjects, direct objects and the above-mentioned indirect objects all fall within the first category, whereas those objects in the accusative (13) or locative case (14) requiring a preposition obviously belong to the prepositional ones.

(13) *Preselit ću se u Amerik-u*
To move (I) will REFL to America-**ACC**
'I am moving to America'

(14) *Živim u Zagreb-u*
(I) live in Zagreb-**LOC**
'I live in Zagreb'

This being said, in order to facilitate future multilingual linking between resources, an attempt was made to keep the template of clause-role components for CROATPAS as adherent as possible to its Italian counterpart. Here is a list of the final clause-role labels used in CROATPAS:

1) **SUBJECT** – nominative and dative subjects
2) **OBJECT** – direct objects in the accusative case and partitive genitives
3) **INDIRECT COMPLEMENT** – indirect objects in the genitive, dative or instrumental case and prepositional objects
4) **ADVERBIAL** – to be used for those obligatory complements expressed by adverbs
5) **CLAUSAL –** for both clausal objects and subjects (sub-labels further specify which)
6) **PREDICATIVE COMPLEMENT** – of both object and subject (sub-labels further specify which)

Since both TPAS and CROATPAS are first and foremost semantic resources, the same verb pattern can contain different syntactic realizations. For instance, the corpus concordances behind the pattern displayed by example (6) contain sentences where the

[5] Please note that, for the purpose of this paper, we limit the morphological glosses to case labels. However, the following examples show a number of typological features worth paying attention to, such as the fact that Croatian is a pro-drop language, it does not have articles and has an SVO word order. Here is a list of the abbreviations that we used: NOM (nominative), GEN (genitive), DAT (dative), ACC (accusative), LOC (locative), INS (instrumental), REFL (reflexive particle), Q (question particle).

SemType [INFORMATION] is assigned to both Objects in the accusative case and Clausal Objects, mostly introduced by Croatian complementizers such as DA, ŠTO (both equivalents of THAT) or KAKO (HOW).

(15) [[HUMAN]$_{\text{NOM}}$] **ČUJE** [[INFORMATION]$_{\text{ACC}}$] | KAKO[INFORMATION]
Na početku ćete čuti upute.| Nisam čuo kako je bilo.
At the start you will hear instructions.|I did not hear how it was.

Last but not least, verbal aspect had also to be taken into account during the set up of CROATPAS. Aspect is a grammatical category which applies to verbs only, offering "different ways of viewing the internal temporal constituency of a situation" (Comrie 1976: 3). Those verbs characterised by an imperfective aspect are able to report about actions while they are being carried out, while others – the perfective ones – focus on the completion of such actions. In some languages, aspect can be expressed through the choice of tense (in Italian, *imperfetto* vs. *passato remoto* or *passato prossimo*) or by means of periphrases (in English, the *-ing* form). On the other hand, Slavic languages such as Croatian present a set of prefixes and suffixes that are able to create so-called aspectual pairs or *vidski parnjaci* from one of the two forms (Barić et al. 1997: 226).

to read : ČITATI – PROČITATI (*imperfective/ perfective*)
to write : PISATI – NAPISATI (*imperfective/ perfective*)
to announce : OBJAVITI – OBJAVLJIVATI (*imperfective/ perfective*)

For each aspectual pair, patterns were extracted keeping the perfective and imperfective variants separate in the resource, as if they were two different verbs. Thus, by comparing the pattern inventories of the two aspects in each pair, we are able to evaluate to what extent aspectual differences influence verb meaning.

5 Related works

As we have already mentioned, CROATPAS is the sister project of the TPAS resource for Italian (Ježek et al. 2014). Both resources follow the CPA methodology (see § 2), which is also applied in the Pattern Dictionary of English Verbs (PDEV, Hanks & Pustejovsky 2005; El Maarouf et al. 2014) and in its Spanish counterpart (PDSV[6]).

Existing reference dictionaries for Croatian are the *e-Glava*[7] online valency dictionary of Croatian verbs (Birtić et al. 2017) and the Croatian Valence Lexicon of Verbs (CROVALLEX[8], Mikelić Preradović et al. 2009). Unlike CROATPAS, *e-Glava* focuses only on 57 psychological verbs, whose meanings have been selected from pre-existing dictionaries and linked to valency patters, which have been manually extracted from various Croatian corpora. Each argument in *e-Glava* is described on a morphological, syntactic and semantic level. As for morphology, the resource takes into account cases, prepositions and sentential realisations such as the complementizers ŠTO, DA, KAKO etc. Ten complement classes are specified at a syntactic level, namely Nominative Complement, Genitive Complement, Dative Complement, Accusative Complement, Instrumental Complement, Prepositional Complement, Adverbial Complement, Predicative Complement, Infinitive Complement and Sentential Complement (Birtić et al. 2017: 45). On a semantic level, the resource takes into account semantic role labelling (Agent, Patient, etc.), but has not yet introduced any hierarchically organised tagset of SemTypes as CROATPAS does.

Another important lexicographic reference work for Croatian is CROVALLEX (Mikelić-Preradović et al. 2009), the first project aiming at building a lexicon of valence frames for Croatian verbs. Its syntactic-semantic classes are taken from VerbNet (Kipper-Schuler 2005), which is based on Levin's verb classes (1993). Once again, morphological information such as case markings and preposition are displayed, as well as semantic roles, but there is no mention of SemTypes. Overall the semantic resource CROATPAS is complementary to existing resources that focus primarily on the morphosyntactic layer.

6 Multilingual pattern linking and computer-assisted language learning

As pointed out by Baisa et al. (2016b), monolingual CPA-based dictionaries offer a unique chance to create multilingual resources by linking corresponding patterns, since they have been created following the same methodology.

[6] PDSV is being compiled at the Pontifical Catholic University of Valparaíso (Chile) and is available online at: http://www.verbario.com (last visited on July 12th 2019). The project is coordinated by Irene Renau.

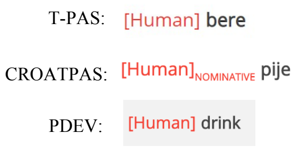
[7] http://valencije.ihjj.hr/page/sto-je-e-glava/1/ (last visited on July 12th 2019)

[8] http://theta.ffzg.hr/crovallex/data/html/generated/alphabet/index.html (last visited on July 12th 2019)

An early attempt of bilingual pattern linking was carried out by Popescu & Ježek (2013), who aligned CPA patterns of English and Italian using examples from the parallel corpus RTE3. Translation pairs were automatically extracted from the corpus and assigned to the corresponding patterns in the source and target language. The study was aimed at testing whether pattern-based translation is more likely to preserve meaning than Google translations, which was proven to be the case. More recently, Baisa et al. (2016a & 2016b) carried out further studies aimed at linking verb patterns from PDEV and its Spanish counterpart (PDSV) via their shared semantic types following both manual procedures and heuristic-based algorithms. Following Baisa, Vonšovský (2016) worked on the automatic linking of PDEV and VerbaLex (Hlavácková 2008), a verb valency lexicon for Czech.

Starting in September 2019, an attempt is being made to cross-linguistically align a sample of 50 verb entries from CROATPAS with their Italian and English counterparts in TPAS and PDEV. We are interested in developing a flexible, semi-automatic, Italian-driven procedure able to disambiguate and link verb patterns across languages by matching their overlapping semantic contexts.

Perfect matches are already clearly foreseeable for verb patterns such as the ones in Figure 2, where both Italian, Croatian and English encode the meaning of "drinking a certain amount of alcoholic beverages" using the SemType [HUMAN] associated with the language-specific equivalent of TO DRINK.



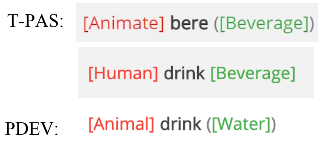
T-PAS: [Human] bere

CROATPAS: [Human]$_{\text{NOMINATIVE}}$ pije

PDEV: [Human] drink

Figure 2 – Perfect pattern matches

In order to be able to link also verb patterns which are not a perfect match, we are developing an algorithm able to recognize pattern similarity by taking into account also hypernym/hyponym relations between SemTypes. Figure 3 provides a fitting example, which shows how different annotation choices can result into the lumping or separation of semantically connected patterns containing hierarchically related SemTypes, such as [ANIMATE] > [HUMAN] & [ANIMAL] or [BEVERAGE] > [WATER].



T-PAS: [Animate] bere ([Beverage])

[Human] drink [Beverage]

PDEV: [Animal] drink ([Water])

Figure 3 – Hierarchically related SemTypes

On the other hand, CROATPAS has also the potential to become an interesting tool for learners and teachers of Croatian as an L2 in computer-assisted language learning (CALL), especially if combined with a user-friendly SKELL-inspired interface (Kilgarriff et al. 2015).

As its creators put it, SKELL (Sketch Engine for Language Learners) is "a stripped-down, non-scary version of Sketch Engine", which grants learners access to:

- a summary of a word's grammatical and collocational behaviour (Word Sketch);
- prototypical example sentences (Good Dictionary Examples) chosen by the GDEX algorithm (Kilgarriff et al. 2008);
- word clouds of similar words, i.e. words that share most collocations with the headword;
- corpus concordance lines

In the case of CROATPAS, displaying Good Dictionary Examples for each of the identified patterns could be a good way to provide real-life context and optional access to more concordance lines could be given to advanced learners. Word clouds displaying the lexical sets populating the SemTypes might also offer an eye-catching opportunity for computer-assisted vocabulary lessons.

At the moment, a resource which is probing these waters is *Woordcombinaties*: a Dutch tool aimed at combining access to collocations, idioms and valency patterns for computer-assisted second language learning and teaching (Colman & Tiberius 2018). This Dutch Collocation, Idiom and Pattern Dictionary focuses on a selection of mid-frequency lexical verbs and aims at offering immediate access to usage patterns from a toolbar, whose search options are: verbs in example sentences, Word Sketches with collocates, pattern-meaning pairs and pragmatic-oriented conversational routines (ibidem. 239). As underlined by the authors, tailor-made examples and Word Sketches can provide a good first impression of an unknown verb, while pattern-meaning pairs are thought for "advanced learners trying to find target collocates or seeking confirmation of their intuitions regarding a collocation” (ibidem. 240).

7 Conclusion

In this paper, we introduced CROATPAS, a corpus-based digital collection of verb valency structures with the addition of semantic type specifications (SemTypes) to each argument slot. The resource relies on Pustejovsky’s Generative Lexicon theory (1995, 1998; Pustejovsky & Ježek 2008) (Section 3) and is made up of four key-components, namely: 1) a representative corpus of contemporary Croatian (hrWac 2.2. RELDI PoS-tagged); 2) a shallow ontology of SemTypes; 3) a methodology for Corpus Pattern Analysis (CPA, Hanks 2004 & 2013); and 4) the adequate corpus tools (Sketch Engine). We discussed the Croatian-specific challenges we faced while building the editor in Section 4, and provided an overview of the existing related works in Section 5. In Section 6, we anticipated the future multilingual linking of verb patterns from CROATPAS, TPAS and PDEV, which could provide a resource to be exploited in NLP, automatic translation and both theoretical and applied cross-linguistic studies. Moreover, CROATPAS could become an interesting tool for computer-assisted language learning (CALL).

References

V. Baisa, S. Može, I. Renau (2016a). Linking Verb Pattern Dictionaries of English and Spanish. Presented at the *5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*. Portorož, Slovenia.

V. Baisa, S. Može, I. Renau (2016b). Multilingual CPA: Linking Verb Patterns Across Languages. In: *Proceedings of the XVII Euralex International Congress*. Tbilisi, Georgia.

E. Barić, M. Lončarić, D. Malić, S. Pavešić, M. Peti, V. Zenčević, M. Znika (1997). *Hrvatska gramatika*. Zagreb: Skolska knjiga.

M. Baroni & A. Kilgarriff (2006). Large Linguistically-Processed Web Corpora for Multiple Languages. In: *Proceedings of the XI Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy.

M. Birtić, I. Brač, S. Runjaić (2017). The Main Features of the e-Glava Online Valency Dictionary. In: *Proceedings of the 5th eLex conference - Electronic lexicography in the 21st century.* Leiden, Netherlands.

S. Cinkova, M. Holub, A. Rambousek, L. Smejkalova (2012). A database of semantic clusters of verb usages. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC ‘12).* Instanbul, Turkey.

L. Colman & C. Tiberius (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In: *Proceedings of the XVIII EURALEX International Congress*. Ljubljana, Slovenia.

B. Comrie (1976). *Aspect: An introduction to the study of verbal aspect and related problems.* Cambridge: Cambridge University Press (6th edition).

T. De Mauro (2016). *Il Nuovo Vocabolario di Base della lingua italiana*. Available at the website: https://www.dropbox.com/s/mkcyo53m15ktbnp/nuovovocabolariodibase.pdf?dl=0 (last visited on July 12th 2019).

I. El Maarouf, J. Bradbury, P. Hanks (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In: *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik, Iceland.

P. Hanks (2004). Corpus Pattern Analysis. In: *Proceedings of the XI Euralex International Congress*. Lorient, France.

P. Hanks (2012). How People use words to make Meanings. Semantic Types meet Valencies. In: A. Bulton and J. Thomas (eds.) *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press.

P. Hanks (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: The MIT Press.

P. Hanks, E. Ježek, D. Kawahara, O. Popescu (2015). Corpus Pattern for Semantic Processing. In: *Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, Beijing, China.

P. Hanks & J. Pustejovsky (2005). A Pattern Dictionary for Natural Language Processing. In: *Revue française de linguistique appliquée*, 10 (2), pp. 63-82.

D. Hlaváckovᴀ́ (2008). *Databáze slovesnchý valenčních rámců VerbaLex (Database of Verb Valency Frames VerbaLex)*, PhD Thesis, Masaryk University, Brno, Czech Republic.

E. Ježek (2016). *The lexicon: An introduction*. Oxford: Oxford University Press.

E. Ježek & V. Quochi (2010). Capturing Coercions in Texts: a First Annotation Exercise. In: *Proceedings* *of the VII conference on International Language Resources and Evaluation (LREC)*. Valletta, Malta.

E. Ježek, B. Magnini, A. Feltracco, A. Bianchini, O. Popescu (2014). T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In: *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.

A. Kilgarriff, M. Husák, K. Mcadam, M. Rundell, P. Rychlý (2008). GDEX : automatically finding good dictionary examples in a corpus. In: *Proceedings of the 13th EURALEX International Congress* (pp. 425–432). Barcelona, Spain.

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovár, J. Michelfeit, P. Rychlý, V. Suchomel (2014). The Sketch Engine: ten years on. In: *Lexicography* 1(1), pp. 7-36.

A. Kilgarriff, F. Marcowitz, S. Smith, J. Thomas (2015). Corpora and Language Learning with the Sketch Engine and SKELL. In: *Revue française de linguistique appliquée*, 20(1), pp. 61-80.

K. Kipper-Schuler (2005). *VerbNet: A broad coverage, comprehensive verb lexicon,* Ph.D. Thesis, University of Pennsylvania, USA.

B. Levin (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: The University of Chicago Press.

N. Mikelić Preradović, D. Boras, S. Kišiček (2009). CROVALLEX: Croatian Verb Valence Lexicon. In: *Proceedings of the 31st International Conference on Information Technology Interfaces*. Zagreb, Croatia.

N. Ljubešić & T. Erjavec (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In: *Text, Speech and Dialogue, Lecture Notes in Computer Science*, Springer.

O. Popescu & E. Ježek (2013), Verbal Phrase Translation, *Tralogy Session 2 - Sense and Machine.* URL: http://lodel.irevues.inist.fr/tralogy/index.php?id=216&format=print (last visited on July 12th 2019).

J. Pustejovsky (1995). *The Generative Lexicon*. Cambridge: The MIT Press.

J. Pustejovsky (1998). The semantics of lexical underspecification. In: *Folia Linguistica* 32.

J. Pustejovsky, P. Hanks, A. Rumshisky (2004). Automated Induction of Sense in Context. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*. Geneva, Switzerland.

J. Pustejovsky & E. Jezek (2008). Semantic Coercion in Language: Beyond Distributional Analysis. In: *Italian Journal of Linguistics*, vol. 20, pp. 181-214.

G. Rehm & S. Hegele (2018), Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs. In: *Proceedings of the XI Language Resources and Evaluation Conference (LREC 2018).* Miyazaki, Japan.

G. Rehm, H. Uszkoreit, I. Dagan, V. Goetcherian, M. U. Dogan, C. Mermer, T. Váradi, S. Kirchmeier-Andersen, G. Stickel, M. Prys Jones, S. Oeter, S. Gramstad (2014). An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In: *Proceedings of the Workshop on Collaboration and Computing for UnderResourced Languages in the Linked OpenData Era (CCURL 2014).* Reykjavik, Iceland.

A. Špikić (2017). *Croato compatto: dizionario croato/italiano e italiano/croato*, Zanichelli: Bologna.

M. Tadić, D. Brozović-Rončević, A. Kapetanović, (2012). *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age.* In: META-NET White Paper Series, G. Rehm & H. Uszkoreit (eds.), Springer: Heidelberg, New York, Dordrecht, London.

J. Vonšovsky (2016). *Automatic Linking of the Valency Lexicons PDEV and VerbaLex* (MA Thesis). URL:http://is.muni.cz/th/359500/fi_m/AutomaticLinking.pdf (last visited on July 12th 2019).