Towards Semantic Role Labeling in Slovene and Croatian

Polona Gantar,* Kristina Štrkalj Despot, Simon Krek,† Nikola Ljubešić‡**

* Department of translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@guest.arnes.si
** Institute for the Croatian Language and Linguistics
Republike Austrije 16, 10000 Zagreb
kdespot@ihjj.hr
† Artificial Intelligence Laboratory, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
simon.krek@ijs.si
‡ Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si

Abstract

In the paper, the semantic role labeling framework is presented, which was developed within the project *Semantic Role Labeling in Slovene and Croatian.* The main goal of the project was the development of an annotated corpus to be used as training data for supervised machine learning systems. In building this framework we follow the path of previous SRL endeavours such as PDT, Vallex, FrameNET, Propbank etc. In compiling the list of semantic roles and their respective formal descriptions, we follow the approach developed by Prague Dependency Treebank, PDT. The paper describes both corpora used for semantic role annotation, as well as tools used in manual annotation tasks. Special attention is directed towards the description of the experimental automatic semantic role labeling based on supervised machine learning methods, and to its possible improvements. A preliminary quantitative analyses is performed for both languages (in terms of verbs range and frequencies, semantic roles, and typical syntactic-semantic patterns for the most frequent verbs).

Označevanje semantičnih vlog za slovenščino in hrvaščino

V prispevku opisujemo model semantičnega označevanja za slovenščino in hrvaščino, ki smo ga razvili v okviru mednarodnega bilateralnega projekta. Osnovni namen projekta je bil izdelati ročno označena korpusa, ki ju bo mogoče uporabiti kot učno množico v sistemih nadzorovanega strojnega učenja za oba jezika. Model sledi dobrim jezikovnim praksam ter široko uveljavljenim modelom na tem področju (PDT, Vallex, FrameNET, Propbank), hkrati pa upošteva značilnosti obeh jezikov kot tudi robustnost semantičnih oznak. V članku opišemo oba učna korpusa in nabor semantičnih oznak ter na kratko povzamemo rezultate poskusnega avtomatskega označevanja s pomočjo nadzorovanega strojnega učenja. V jedrnem delu prispevka opišemo prve rezultate kvantitativnih analiz za oba jezika, in sicer z vidika zastopanosti glagolov, semantičnih oznak in tipičnih pomensko-skladenjskih vzorcev za najfrekventnejše glagole.

1. Introduction

Semantic Role Labeling (SRL) within natural language processing refers to the process of detecting and assigning semantic roles to semantic arguments determined by the predicate or verb of a sentence. This means that in the sentence *My parents gave me a weird name*, the verb *to give* should be recognized as the predicate with three arguments: the one who deliberately performs the action or the agent (*parents*), the one who is the recipient or the experiencer of the event (*me*), and the one that undergoes the action or the patient/theme of the action (*name*). The analysis of semantic roles (both of the arguments and adjuncts) is important both within theoretical linguistics and within applied linguistics in compiling semantic lexicons and valency dictionaries. From the point of view of language technologies, the task of semantic role labeling is important within the development of the information extraction systems, question answering systems, improving syntactic parsing systems, in machine translation tasks etc. (Shen in Lapata, 2007; Christensen et al., 2011). In comparison with syntactic trees, semantic role labeling requires higher level of abstraction, and it is a very important step towards the understanding of the meaning of a sentence. This is why SRL plays a major role in natural language processing. For instance, in the sentence *A weird name was given to me by my parents*, the morphosyntactic representation of the sentence is different than in the sentence mentioned earlier. However, semantic roles are the same in both sentences.

A comprehensive comparative analysis performed within META-NET white book series (Krek et al., 2012) has shown that both Slovene and Croatian may be considered as under-resourced languages in terms of language technologies, especially in the area of machine readable semantic resources and advanced tools for the processing of those resources.

Therefore, SRL will improve the existing levels of linguistic annotation of both Slovene and Croatian training corpora. With close cognate[1] languages it is advisable and beneficial to use similar principles and annotation schemes in the same natural language processing tasks.

Therefore, a project *Semantic Role Labeling in Slovene and Croatian* was conducted. The aim of the project was to build a semantic role labeling system which will be added

[1] Both languages in question belong to South Slavic branch of Slavic language family.

to the existing syntactic dependencies in both Slovene and Croatian training corpora used hitherto for machine learning algorithms. The core project tasks included: 1) development of the common Slovene-Croatian semantic annotation scheme and the creation of the list of semantic role labels based on the existing resources for other languages; 2) compiling the instructions for annotation; 3) manual annotation of the sample parts of both learning corpora using compatible tags. This served as the basis for the automatic annotation experiments using supervised machine learning methods, performed later on both corpora.

In the paper, we will present the resulting semantic role labeling framework in detail. The framework follows the path of similar previous SRL endeavours such as PDT, Vallex, FrameNET, Propbank, Crovallex etc. (see Krek et al., 2016). The paper describes both corpora used for semantic role annotation, as well as tools used in manual annotation tasks. Special attention is directed towards the description of the data obtained from the experimental automatic semantic role labeling based on supervised machine learning methods, and to its possible improvements. A preliminary quantitative analysis is performed for both languages (in terms of verbs range and frequencies, semantic roles, and typical syntactic-semantic patterns for the most frequent verbs).

2. Semantic Role Labeling framework for Slovene and Croatian

In compiling the list of semantic roles and their respective formal descriptions, we follow the approach developed by Prague Dependency Treebank, PDT (Mikulová et al., 2005), in which verbs or predicates determine arguments and adjuncts (usually specifying circumstances: time, location etc.). In addition, multi- word predicate role can be specified. (Table 1).

In the framework which was developed for the annotation of the Slovene and Croatian corpus, in addition to PDT, we have consulted Valency Lexicon of Czech Verbs (Vallex), semantic role labeling within Croatian Dependency Treebank (SRL tagset compiled by Filko et al. 2012), and Crovallex (Croatian version of Czech Vallex) which contains 1740 verbs selected from the Croatian frequency dictionary (Mikelić Preradović et al., 2009).

Our final SRL tagset (Table 1) contains 25 semantic labels (5 of those are arguments, 17 adjuncts, and 3 labels for multi-word predicates). The concept of obligatoriness or “coreness” was not used in the framework as compatible semantic resources (e.g. valency lexicons or FrameNets with a defined concept of obligatoriness) for both languages are not available at the moment.

SLO/CRO		**PDT**	
agent	ACT	actor	ACT
patient	PAT	patient	PAT
recipient	REC	addressee	ADDR
		benefactor	BEN
origin	ORIG	origo	ORIG
		inheritence	HER
result	RESLT	effect	EFF
location	LOC	direction	DIR2
		locative	LOC
source (location)	SOURCE	direction	DIR1
goal (location)	GOAL	direction	DIR3
event	EVENT		
time	TIME	temporal	TWHEN
		temporal	TPAR
		temporal	TFRWH
		temporal	TOWH
duration	DUR	temporal	TFHL
		temporal	THL
		temporal	TSIN
		temporal	TTILL
frequency	FREQ	temporal	THO
aim	AIM	aim	AIM
		intent	INTT
cause	CAUSE	cause	CAUS
contradiction	CONTR	contradiction	CONTRD
		concession	CNCS
condition	COND	condition	COND
regard	REG	regard	REG
		criterion	CRIT
		comparison	CPR
accompaniment	ACMP	accompaniment	ACMP
restriction	RESTR	restriction	RESTR
manner	MANN	manner	MANN
		result	RESL
means	MEANS	means	MEANS
quantification	QUANT	difference	DIFF
		extent	EXT
multi-word predicate	MWPRED		
modal	MODAL		
phraseological unit	PHRAS	dependant part of phraseme	DPHR

Table 1: SRL Tagset in SLO/CRO in comparison with PDT system.

3. Corpora and Tools for Annotation

On the Slovene side, the SSJ500k 2.0 (Krek et al., 2015) corpus was used for manual annotation of semantic roles. The corpus contains 500,293 words (27,829 sentences) sampled from the FidaPLUS corpus (Arhar Holdt and Gorjanc, 2007). The whole corpus is manually annotated on morphosyntactic level (Grčar et al., 2012), and partly on syntactic level (Dobrovoljc et al., 2012). Named entities and multi-word expressions are also identified (Gantar et al., 2017). The total of 5,491 sentences were annotated with semantic roles, with the first 500 sentences used for test annotation by four annotators. The second phase included automatic annotation (see Chapt. 3.1) of the remaining 4,991 sentences and their manual check by 5 annotators. These represent the basis for the quantitative analysis of the Slovene training corpus.

For the Croatian language, we used the SETimes.HR part of the hr500k corpus (Ljubešić et al., 2018), which is based on a sample of the Croatian part of the SETimes parallel corpus. It contains 3,757 sentences manually lemmatized and morphosyntactically tagged (Agić et al., 2013), and annotated for syntactic dependencies using the Universal Dependencies formalism (Agić and Ljubešić, 2015). Within this project, these sentences were being manually semantically annotated by 2 annotators. This then served as the resource for automatic labeling and quantitative analysis.

3.1. Automatic Semantic Role Labeling

Both annotated corpora were split in training and test data in a 80:20 fashion. This data split is available for each of the languages at https://github.com/clarinsi/bilateral-srl/tree/master/data.

Publishing the specific data split publicly has the goal of fostering comparing various tools on both languages and identifying that or those that perform best, or with the minimum memory and time footprint.

Currently the well-known baseline mate-tools semantic role labeler (Björkelund et al. 2009) was benchmarked on the data with the per-label F1 metric reported in Table 2. The weighted F1 score for all classes for Croatian was 0.72, while for Slovene it was 0.75. The parser is available from https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/mate-tools/srl-4.31.tgz, and it was used without any modifications, using the German feature set.

Label	Croatian	Slovene
PAT	0.81	0.88
ACT	0.91	0.94
RESLT	0.83	0.80
TIME	0.65	0.62
REC	0.78	0.74
MODAL	0.94	0.90
MANN	0.45	0.76
LOC	0.56	0.59
DUR	0.64	0.50
ORIG	0.65	0.24
CAUSE	0.14	0.35
REG	0.43	0.34
AIM	0.47	0.20
GOAL	0.38	0.53
QUANT	0.54	0.62
MWPRED	0.72	0.91
EVENT	0.68	0.29
ACMP	0.80	0.08
MEANS	0.44	0.64
FREQ	0.50	0.59
CONTR	0.21	0.14
COND	0.59	0.46
PHRAS	0.11	0.31
SOURCE	0.29	0.37
REST	0.0	0.0

Table 2: Results (F1) of the experiments on automatic labeling of Croatian and Slovene with mate-tools for each label.

The data on both languages are quite similar, with F1 metrics corresponding to the frequency of each phenomenon. More concretely, on the Croatian dataset, the Pearson correlation between frequency and F1 is 0.517 with a p-value of 0.008, while on the Slovene dataset the same correlation coefficient is 0.611 with a p-value of 0.001. We can conclude that both correlation coefficients are strong and statistically highly significant

4. Quantitative analyses

In the next chapters, the preliminary quantitative analysis of both corpora is presented from the point of view of verb frequencies, semantic roles, and syntactic-semantic patterns that are recognized in the corpus as being stable and typical for individual verbs (here only for the most frequent verbs).

4.1. Verbs representation in both corpora

The Slovene SRL-annotated corpus contains 15,988 verbal tokens with 1,953 lemmas. The percentage of verbal lemmas appearing only once in the corpus is 47,5.

The Croatian SRL-annotated corpus contains 12,605 verbal tokens with 1,094 lemmas. The percentage of verbal lemmas appearing only once in the corpus is 40.8.

As expected, most frequent in both corpora are verbs with broad meaning spectrum such as *biti* ‘to be’, *imeti/imati* ‘to have’, *dobiti* ‘to get’; modal verbs: *morati* ‘must’, *moči/moći* ‘can’, *hoteti/htjeti* ‘will’, *želeti/željeti* ‘want’, and verbs of communication *reči/reći* ‘to say’, *povedati/kazati* ‘to tell’. Significantly higher frequency of the verbs of communication in the Croatian corpus (*kazati, izjaviti, reći, priopćiti, navoditi* = ‘to tell, say, state etc.’) is the result of the fact that SETimes.HR corpus consists only of news texts.

The list of verb lemmas with the minimum frequency of 50 in Slovene and Croatian corpora are in Table 3.

SSJ500k		SETimes.HR	
biti	7203	**biti**	4969
imeti	333	**htjeti**	670
morati	178	**kazati**	276
iti	114	izjaviti	210
vedeti	95	**moći**	195
dobiti	83	**imati**	163
moči	83	**reći**	160
začeti	80	trebati	146
videti	75	**morati**	117
reči	74	**željeti**	65
priti	72	očekivati	62
povedati	72	**dobiti**	57
hoteti	69	**postati**	57
želeti	59	postojati	56
postati	54	priopćiti	54
govoriti	51	predstavljati	53
misliti	50	navoditi	50

Table 3: Verbs with frequency f>=50 in SSJ500k and SETimes.HR. Verbs that are present in both corpora are indicated in bold.

Further qualitative analysis included the most frequent verbs (Table 3) and arguments (Figure 1). In case of arguments, we considered their presence in various patterns and their frequency in patterns. Individual verbs were taken as the basis for pattern formulation, however, polysemy (in case of polysemous verbs) was not taken into account. The reason for this is non-existence of compatible valency lexicons in Slovene and Croatian, and the size of the annotated corpora which cover only a limited set of senses per verb. This will be addressed in future versions of SRL annotations when this type of resources will be available.

4.2. Semantic roles representation in both corpora

All 25 semantic labels proposed in our framework are found in both training corpora. As can be observed from the Figure 1, the most frequent semantic roles in both corpora are argument roles of PAT, ACT and RESLT. They are followed by adjunct roles of TIME, MANN, and LOC (the last two being significantly more frequent in the Slovene corpus). In addition to these, other notable differences include significantly higher frequency of patients (PAT) and recipients (REC) in the Slovene corpus. On the other hand, the frequency of agent roles is extremely balanced in both corpora.

A more detailed comparative analysis could explain weather these differences are the result of differences in the corpora design (the Slovene and the Croatian corpora differ in genre representation - the Croatian one containing primarily news texts while the Slovene one being balanced in terms of genre representation). However, different genre representation in corpora certainly has the effect on the higher frequency of communication semantic group of verbs in the Croatian data.

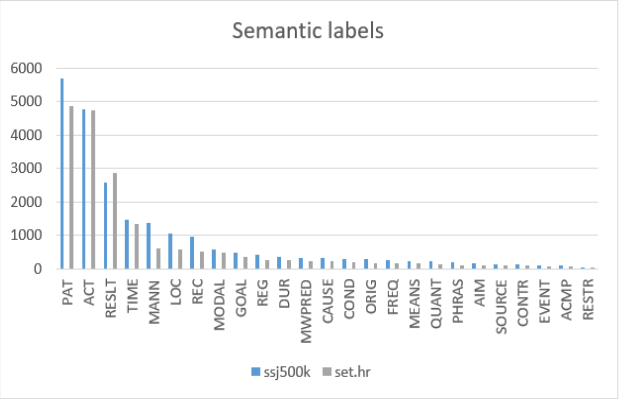

Figure 1: Semantic roles (labels) in Slovene (SSJ500k) and Croatian (SETimes.HR) training corpus.

Fquency of verbs in both corpora is relevant in relation to frequency of arguments in their patterns. Semantic roles with 50 or more hits in patterns are similar in both languages in case of verbs with similar basic meaning (in Table 4 and 5 indicated in bold).

biti	ACT	PAT	RESLT	TIME	MANN
imeti	ACT	PAT			
iti	ACT				
dobiti		PAT			
videti		PAT			
vedeti			RESLT		
postati			RESLT		

Table 4: Most frequent label (f>=50) per verb in Slovene SSJ500k

biti	ACT	PAT	RESLT	TIME
kazati	ACT		RESLT	
izjaviti	ACT		RESLT	TIME
reći	ACT		RESLT	
moći	ACT			
trebati	ACT			
imati	ACT	PAT		
uključivati		PAT		
predstavljati		PAT		
dobiti		PAT		
postati			RESLT	
priopćiti			RESLT	

Table 5: Most frequent label (f>=50) per verb in SETimes.HR

A verb *to be*, due to its broad semantics, is able to take on all of the semantic roles in both languages. Among the most frequent verbs, there are a few other such verbs with obligatory semantic roles (arguments): *imeti/imati* 'to have' (WHO has WHAT), *dobiti* 'to get' (WHO gets WHAT), and *postati* 'become' (WHO becomes WHO/WHAT).

4.3. Syntactic-semantic patterns

From both corpora, we have extracted stable syntactic-semantic patterns characteristic for each individual verb. Those patterns are similar in both languages despite the differences in the corpus design. Here, we will list those patterns (together with the example of their exact linguistic realization from the corpus) for the most frequent verbs in both corpora. To make the formalizations of these patterns more readable, we use "Who did What to Whom, and How, When and Where?" form (ACT = Who, PAT = What, RESLT=Who/What, LOC = Where etc.). Semantic tags are being put in the brackets next to their respective pronouns. The first part of the pattern represents its stable section which includes arguments that are typical for the given verb. In relation to (non-)obligatory nature of arguments, it has to be mentioned that patterns do not include arguments that are "obligatory" but are not explicitly present, e.g. agents (ACT) included in finite verbal forms, as exemplified in case of verbs *vedeti, začeti, videti, reči* etc. Since verb *biti* (to be) is found in combination with all arguments, this pattern was omitted in the analysis of both corpora.

As is the case with PropBank, our framework is also, at this stage, more focused on literal meaning and we did not clearly mark metaphorical usages.

4.3.1 SSJ500k

Slovene training corpus contains relatively stable patterns in case of verbs *imeti, morati, iti, vedeti, dobiti, moči* etc. (*potrebno/treba je* are also in this category) which appear in the corpus more than 70 times:

'to have' *imeti* (333)

- WHO (ACT) has WHAT (PAT 316) [for WHOM (REC), from whom (ORIG), where (LOC), when (TIME) ...]: *Na zadnji hrbtni bodici ima veliko črno piko.*

'must' *morati* (178)

- WHO (ACT) must INF (MODAL): *Država bi morala plačati stroške presoje vplivov na okolje.*

'to go' *iti* (114)

- WHO (ACT) goes WHERE (GOAL) [how

(MANN), when (TIME), under what conditions (COND) …]: *Šel sem prvič k vedeževalki.*
- to go (PHRAS 11): *Zgodba mi ni in ni šla iz glave.*
- to go SUPINE (MWPRED): *Verjetno bom šla smučat na Krvavec.*

‘to know’ *vedeti* (95)
- to know WHAT (RESLT) [how (MANN) …]: *Je pa treba nekaj jasno vedeti.*

‘to get’ *dobiti* (83)
- WHO (ACT) gets WHAT (PAT) [from whom (ORIG), in regard to (REG), with what (MEANS), when (TIME), under what conditions (COND) …]: *Mala je dobila ime po Prometeju.*

‘can’ *moči* (83)
- WHO (ACT) can INF (MODAL): *Ne moremo ga spregledati.*

‘to start’ *začeti* (80)
- WHO (ACT) starts WHAT (PAT) [how often (FREQ), when (TIME) …]: *Razpravo o tem je začel parlamentarni odbor.*
- to start INF (MWPRED): *Najprej začne pripravljati sladice.*

‘to see’ *videti* (75)
- WHO (ACT) sees WHAT (PAT) [when (TIME), in regard to what (REG), from where (SOURCE) …]: *Tukaj so jo zadnjič videli.*

‘to say’ *reči* (74)
- WHO (ACT) says TO WHOM (REC) WHAT (RESLT) [when (TIME) …]: *Neka psihologinja mi je rekla, da moram živeti le zase.*

‘to come’ *priti* (72)
- WHO (ACT) comes [to what (RESLT), when (TIME), where (GOAL), by what means (MEANS), why (CAUSE), under what conditions (COND) …]: *Na kongres v Ljubljano je prišlo več kot 500 gostov.*
- to come (PHRAS): *Vse to je prišlo na dan.*

Also, verb *iti* needs to be explained, with its pattern with the prepositional phrase *gre za* + WHO/WHAT (ENG: it is about). In this case the semantic role chosen for the argument expressed with WHO/WHAT was agent and not patient: *gre za vprašanje/preteklost/rešitev* etc. (ACT) (ENG: it’s about the question/past/solution). If the verb in the same pattern is used for expressing motion, e. g. *iti v Evropo/samostan/desno* (GOAL) (ENG: to go to Europe/monastery/the right) the agent is not necessarily present. The verb *iti* and its counterpart *priti* are also somewhat special in the sense that they form phraseological units such as *ne iti v račun* (‘not being able to comprehend’), *ne iti iz glave* (‘not being able to forget’), *iti na živce* (‘to make nervous’), *priti v poštev* (‘to (be able to) be considered’) with the label PHRAS.

4.3.2. SETimes.HR

From the Croatian training corpus, we have recognized and extracted fixed and stable syntactic-semantic patterns in case of verbs that appear in the corpus more than 50 times (*htjeti, kazati, moći, imati, trebati* etc.).

‘to want’ *htjeti* (670), ***željeti*** (65)
- WHO (ACT) wants WHAT (PAT) [for WHOM (REC), from WHOM (ORIG)...]: *Oni žele autonomiju sjevera, a za druge enklave žele takozvani Ahtisaari plus.*
- WHO (ACT) wants INF (MODAL): (WHAT) (PAT): *Mnoge žrtve ne žele podnijeti tužbu.*

‘to tell, say’ *kazati* (276), ***izjaviti*** (210), ***reći*** (160)
- WHO (ACT) says WHAT (RESLT) to WHOM (REC) about WHAT (PAT) [WHERE (LOC), WHEN (TIME)]: “*U suprotnom ćemo biti neozbiljni političari”, rekao je Lagumdžija novinarima u Beogradu nakon sastanka s Jeremićem 14. ožujka.*

‘can’ *moći* (195)
- WHO (ACT) can INF (MODAL) WHAT (PAT): *Privatizacija je mogla donijeti bolje usluge.*

‘to have’ *imati* (163)
- WHO (ACT) has WHAT (PAT) [WHEN (TIME) for WHOM (REC), from WHOM/WHAT (ORIG)...]: *Moldavija sada ima novog predsjednika.*
- imati u vidu (PHRAS): *Imajući u vidu nesuradnju Beograda s Haaškim tribunalom ...*
- [(WHO) (ACT)] imati za cilj (PHRAS) WHAT (PAT): *Reforme za cilj imaju stavljanje oružja pod nadzor.*

‘to need’ *trebati* (146)
- WHO (ACT) needs INF (MODAL) WHAT (PAT) [to WHOM (REC)]: *Mi trebamo dati potporu Jeremiću.*

‘must’ *morati* (117)
- WHO (ACT) must INF (MODAL): *Čelnici moraju voditi.*

‘to expect’ *očekivati* (62)
- WHO (ACT) expects WHAT (PAT): *Katastarski dužnosnici očekuju registraciju oko 6,7 milijuna katastarskih čestica.*

‘to get’ *dobiti* (57)
- WHO (ACT) gets WHAT (PAT): *Manjinske zaklade dobit će naknadu za imovinu.*

5. Summary and Conclusions

In the paper, the data obtained from the experimental automatic semantic role labeling based on supervised machine learning methods, and the preliminary quantitative analyses of Slovene and Croatian training corpora (in terms of verbs range and frequencies, semantic roles, and typical syntactic-semantic patterns for the most frequent verbs) are presented.

The data for both languages are quite similar from all the above perspectives, despite the differences in corpora design.

From the preliminary analysis of the data, it seems that the SRL framework that was being developed within this bilateral project is suitable for semantic role labeling tasks in both languages. Moreover, the framework has been successfully implemented to serve as the solid base for the automatic SRL (using supervised machine learning methods).

Having a common framework for semantic annotation of cognate languages (Slovene and Croatian) was proved to be advantageous in terms of saving time and resources. Moreover, developing and applying a common framework was very beneficial from the perspective of mutual evaluation and corrections as well. This framework is also a solid base for future more detailed comparative semantic analyses.

Building a corpus with SRL annotations is an ongoing work and both corpora will be upgraded in the future. Upgrades will include the increase in size, calculation of inter-annotator agreement and segmentation of patterns according verb senses (when compatible semantic resources for both languages are available).

6. Acknowledgments

The paper was prepared as part of two projects, Nova slovnica sodobne standardne slovenščine: viri in metode (New grammar of contemporary standard Slovene: sources and methods, ARRS J6-8256) and by Slovene-Croatian bilateral government program. We would like to thank annotators of both corpora (in alphabetical order): Lucija Jezeršek, Taja Kuzman, Dafne Marko, Ivan Pandžić, Iza Škrjanec and Anja Zajc.

7. References

Špela Arhar Holdt and Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2), 95–110.

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, pages 48–57. Sofia, Bulgaria, Association for Computational Linguistics.

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of the Workshop on Balto-Slavic Natural Language Processing*, pages 1–8.

Collin F. Backer, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, pages 86–90. Montreal, Canada.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of The Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, Boulder, June 4–5, pages 43–48.

Christensen, Janara, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An Analysis of Open Information Extraction based on Semantic Role Labeling. *International Conference on Knowledge Capture* (KCAP), pages 113–120. Banff, Alberta, Canada.

Kaja Dobrovoljc, Simon Krek, and Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. In *Zbornik Osme konference Jezikovne tehnologije*, pages 42–47. Ljubljana, Institut Jožef Stefan.

Matea Filko, Daša Farkaš, and Danijela Merkler. 2012. SRL Tagset for Croatian. Institute of Linguistics, Faculty of Humanities and Social Sciences, Zagreb. http://hobs.ffzg.hr/static/docs/SRL_tagset.pdf.

Polona Gantar, Simon Krek, and Taja Kuzman. 2017. Verbal multiword expressions in Slovene. *Europhras 2017*, pages 247–259. Springer.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. *Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik.* In *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana, Institut Jožef Stefan.

Simon Krek. 2012. *Slovenski jezik v digitalni dobi*. Berlin, Heilderberg, Springer Verlag.

Simon Krek, Polona Gantar, Kaja Dobrovoljc, and Iza Škrjanec. 2016. Označevanje udeleženskih vlog v učnem korpusu za slovenščino. In *Proceedings of the Conference on Language Technologies & Digital Humanities*, Faculty of Arts, pages 106–110. University of Ljubljana.

Krek, Simon et al. 2015. Training corpus ssj500k 1.4, *Slovenian language resource repository* CLARIN.SI, http://hdl.handle.net/11356/1052.

Nives Mikelić Preradović, Damir Boras, and Sanja Kišiček. 2009. CROVALLEX: Croatian Verb Valence Lexicon. In *Proceedings of the 31st International Conference on Information Technology Interfaces*, pages 533–538.

Marie Mikulová et al. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. *Annotation manual. Technical Report 30*, pages 5–11.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1): 71–106.

Dan Shen and Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*, pages 12–21. Prague.