Universal Dependencies for Croatian (that Work for Serbian, too)

Željko Agić
University of Copenhagen, Denmark
zeljko.agic@hum.ku.dk

Nikola Ljubešić
University of Zagreb, Croatia
nljubesi@ffzg.hr

Abstract

We introduce a new dependency treebank for Croatian within the Universal Dependencies framework. We construct it on top of the SETIMES.HR corpus, augmenting the resource by additional part-of-speech and dependency-syntactic annotation layers adherent to the framework guidelines. In this contribution, we outline the treebank design choices, and we use the resource to benchmark dependency parsing of Croatian and Serbian. We also experiment with cross-lingual transfer parsing into the two languages, and we make all resources freely available.

1 Introduction

In dependency parsing, the top-performing approaches require supervision in the form of manually annotated corpora. Dependency treebanks are costly to develop, and they typically implement different annotation schemes across languages, i.e., they are not homogenous with respect to the underlying syntactic theories (Abeillé, 2003). Today we know this hinders research in cross-lingual parsing (McDonald et al., 2011), and subsequently the enablement of language technology for under-resourced languages.

The Universal Dependencies (UD) (Nivre et al., 2015) project[1] aims at addressing the issue by providing homogenous dependency treebanks. The treebanks feature uniform representations of parts of speech (POS), morphological features, and syntactic annotations across 18 languages in the current release (Agić et al., 2015).[2] The POS tagset is a superset of Petrov et al. (2012), while the dependency trees draw from the universal Stanford dependencies of de Marneffe et al. (2014). The intricacies of UD are well beyond the scope of our contribution. Instead, we spotlight the parsing and cross-lingual processing of two South East European (SEE) under-resourced languages (Uszkoreit and Rehm, 2012).

In their pivotal contribution to cross-lingual parsing, McDonald et al. (2013) reveal the twofold benefits of uniform representations, as they i) enable more exact evaluation of dependency parsers, and ii) facilitate typologically motivated transfer of dependency parsers to under-resourced languages with improved accuracies. In short, their research indicates that enabling POS tagging and dependency parsing for, e.g., Macedonian would largely benefit should a treebank for a similar language—say, Croatian—exist within an uniform representations framework such as UD.

This work opened up a cross-lingual parsing research avenue that addresses issues such as multi-source transfer, in which multiple source treebanks are combined to improve target language parsing (McDonald et al., 2011), or annotation projection, in which the trees are transferred via parallel corpora and parsers trained on the projections (Tiedemann, 2014). Apart from dependency parsing, this line of work also includes the developments in cross-lingual POS tagging, mainly drawing from the work of Das and Petrov (2011), even if seeded much earlier through the seminal work of Yarowsky et al. (2001). Most of this work, however, does not include the under-resourced SEE languages, and thus we stress that topic in particular in our paper.

Contributions. We focus on dependency parsing of two under-resourced South Slavic languages, Croatian and Serbian, and its implications on cross-lingual parsing of related languages. We list the following contributions: i) a novel, UD-conformant dependency treebank for Croatian, ac-

[1] http://universaldependencies.github.io/docs/

[2] http://hdl.handle.net/11234/LRT-1478

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 1–8,
Hissar, Bulgaria, 10–11 September 2015.

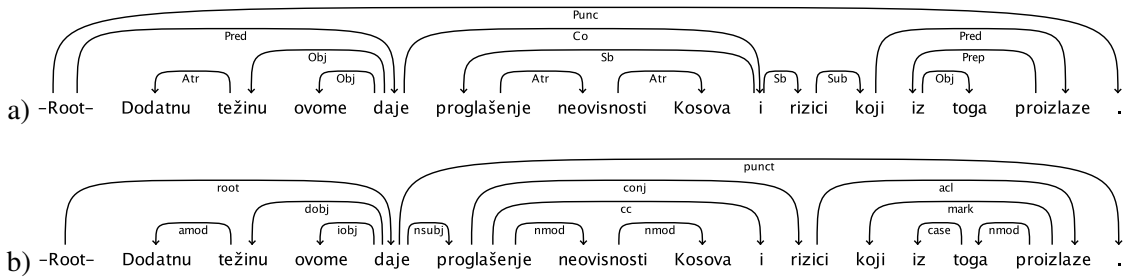

Figure 1: An example sentence from the treebank (training set, #143), with a) SETIMES.HR, and b) UD annotations. Gloss: *Added weight to-this gives the-proclamation of-independence of-Kosovo and the-risks that from it arise.*

companied by cross-domain test sets for Croatian and Serbian, ii) a set of experiments with parsing the two languages within the UD framework, and iii) cross-lingual parsing experiments targeting Croatian and Serbian by source models from two sets of 10 treebanks. We make our datasets available under free-culture licensing.[3]

2 Treebank

UD requires adherence to POS tagset, dependency attachment, and edge labeling guidelines, as well as to the universal morphological feature specifications, the inclusion of which is at this point not mandatory. We provide an UD treebank for Croatian, implementing all the annotation layers.

2.1 Text

Our treebank is built on top of an existing Croatian corpus, the SETIMES.HR dependency treebank (Agić and Ljubešić, 2014). We apply the UD annotation layers on top of its training and testing sets. The sample amounts to 3,557 training sentences of newspaper text, and another 200 development sentences from the same source, which sums up to the 3,757 sentences of the original SETIMES.HR corpus. The training sets are available for Croatian and Serbian, from newswire and Wikipedia, equaling $4 \times 100 = 400$ sentences.

In summary, we take the Croatian text from the SETIMES.HR treebank as a basis for building the Croatian UD treebank, and we include its training, development and test sets in the process. SETIMES.HR also provides Serbian test sets, so we include those as well. As a result, we provide a multi-layered linguistic resource for Croatian and Serbian, offerring two layers of morphological and syntactic annotations on top of the same text. While the usefulness of this particular approach in contrast to opting for an entirely different text sample could be argued, our decision was motivated by i) facilitating empirical comparability across different annotation schemes, and by ii) the line of work by Johansson (2013) with combining diverse treebanks for improved dependency parsing, which we wish to explore in future work focusing on sharing parsers between closely related languages.

2.2 Morphology

SETIMES.HR implements the Multext East version 4 morphosyntactic tagset (MTE4) (Erjavec, 2012). We manually convert it to UD's universal POS tags (UPOS) and universal morphological features, and we make the mapping available with the treebank. Out of the 17 UPOS tags, 14 are used in our treebank, leaving out determiners (DET), interjections (INTJ), and symbols (SYM) as no respective tokens of these types were instantiated in the treebank text. We cast all MTE4 abbreviations into the appropriate UPOS tags—predominantly as nouns, but sometimes also as adverbs such as the Croatian equivalent of "e.g." ("npr.")—by observing the sentence contexts. We also map all the MTE4 morphology into the universal feature set, which accounts for a total of 540 morphosyntactic tags, compared to the 662 in the original dataset, as certain MTE4 features are currently not present in the UD specification. We closely adhere to UD, i.e., we do not introduce any language-specific features at this point.

2.3 Syntax

The annotation for syntactic dependencies was conducted manually by four expert annotators. We decided in favor of manual annotation over implementing an automatic conversion from SE-

[3] https://github.com/ffnlp/sethr

Syntactic tag	%	Gloss	Syntactic tag	%	Gloss
acl	1.89	adjectival clause	expl	0.00	expletive
advcl	0.70	adverbial clause modifier	foreign	0.01	foreign words
advmod	2.12	adverbial modifier	goeswith	0.08	goes with
amod	8.34	adjectival modifier	iobj	0.22	indirect object
appos	1.69	appositional modifier	list	0.00	list
aux	4.35	auxiliary	mark	3.59	marker
auxpass	0.71	passive auxiliary	mwe	0.32	multi-word expression
case	9.80	case marking	name	1.56	name
cc	3.09	coordinating conjunction	neg	0.30	negation modifier
ccomp	1.03	clausal complement	nmod	17.05	nominal modifier
compound	3.02	compound	nsubj	5.97	nominal subject
conj	3.80	conjunct	nsubjpass	0.65	passive nominal subject
cop	1.41	copula	nummod	2.05	numeric modifier
csubj	0.12	clausal subject	parataxis	1.47	parataxis
csubjpass	0.03	clausal passive subject	punct	12.86	punctuation
dep	0.01	unspecified dependency	remnant	0.14	remnant in ellipsis
det	0.98	determiner	root	4.51	root
discourse	0.71	discourse element	vocative	0.00	vocative
dislocated	0.01	dislocated elements	xcomp	1.50	open clausal complement
dobj	3.92	direct object			

Table 1: Syntactic tags in Croatian UD, sorted alphabetically, and listed together with their relative frequencies and short glosses. The frequencies are calculated for Croatian only, and for the entire collection (train, dev, test). The syntactic tags are further explained in the UD documentation: `http://universaldependencies.github.io/docs/u/dep/all.html`.

TIMES.HR to provide Croatian UD with a clean, unbiased start, contrasting the manual creation experience of McDonald et al. (2013) to the one of automatic conversions within the HamleDT project of Zeman et al. (2014).

As with morphology, we use only the universal dependency relations, without introducing language-specific dependency relations. We apply 39 out of 40 universal relations, leaving out only a single speech-specific function (reparandum). We list all the relations with their relative frequencies in Table 1. The annotators strictly adhered to the UD attachment rules, which focus on the primacy of content words in governing dependency relations, which is different from all the existing annotations of Croatian syntax (Agić and Merkler, 2013). Once again, as a general discussion on UD is well beyond the scope of our contribution, we refer the reader to the official UD documentation for all matters relating to the formalism itself. Instead, we focus on a brief comparison of Croatian UD and SETIMES.HR regarding their dependency annotations.

The two schemes apparently differ both in the sets of dependency relations, and in the attachment rules. For the most part, the 15 syntactic tags of SETIMES.HR are generalizations of the 39 Croatian UD concepts. As for the attachment rules, we exemplify some of the differences in Figure 1. First and foremost, there are apparent differences in the treatment of coordination and subordination. In SETIMES.HR, coordinated subjects ("proglašenje" and "rizici") are governed by the coordinator ("i"), while in UD, the first encountered subject ("proglašenje") is assigned the subject role, and the remaining two coordination members are attached to it as siblings with distinct labels. Subordinate clauses are governed by subordinating conjunctions in SETIMES.HR, and in UD, the conjunction ("koji") is attached to the clause predicate ("proizlaze"). A similar rule applies to prepositional phrases ("iz toga"). There are also minor differences in the treatment of genitive complements.

We also look into the non-projectivity of the two syntactic annotation layers. We note from the work by Agić et al. (2013b) that approximately 20% of sentences are non-projective in a Prague-

		Croatian				Serbian					
		NEWS		WIKI		NEWS		WIKI		OVERALL	
Treebank	Features	*UAS*	*LAS*	*UAS*	*LAS*	*UAS*	*LAS*	*UAS*	*LAS*	*UAS*	*LAS*
SETIMES.HR	MTE4 POS	82.2	76.3	77.1	67.9	80.8	74.0	79.8	71.1	80.0	72.3
	+ MTE4 FEATS	84.3	79.2	80.7	73.7	83.0	77.8	82.6	74.7	82.7	76.4
Croatian UD	UPOS	84.8	77.9	80.8	72.4	82.4	75.8	82.1	75.2	82.5	75.3
	+ UPOS FEATS	86.9	81.5	84.5	77.3	86.0	81.5	83.7	77.9	85.3	79.6

Table 2: Parsing accuracy on Croatian and Serbian test sets for the lexicalized models trained on the two Croatian treebanks. Overall scores are highlighted.

style treebank of Croatian (HOBS) (Tadić, 2007). We observe that 10.1% of all sentences are non-projective in SETIMES.HR, while the UD syntax further lowers this figure to only 7.6%. This bears relevance in dependency parsing, as long-distance non-projective relations are more difficult to retrieve by dependency parsers. To some extent, it also reflects the scheme-dependent properties of languages, as it is hard to argue about the exact amount of non-projectivity in Croatian beyond simply confirming its existence given these three distinct figures.

3 Experiments

We conduct two sets of experiments. The first one features monolingual parsing of Croatian and the transfer, albeit trivial, of Croatian parsers to Serbian as a target language, while in the second one, we transfer delexicalized parsers from a number of well-resourced languages to Croatian and Serbian as targets in a cross-lingual parsing scenario.

3.1 Setup

Parser. In all our test runs, we use the graph-based parser of Bohnet (2010).[4] It trains and parses very fast, and it records top-level performance across a number of morphologically rich languages (Seddah et al., 2013). Other than that, it natively handles non-projective structures, which is an important feature for languages such as Croatian and Serbian, and treebanks exhibiting non-projectivity in general. We evaluate using standard metrics, i.e., labeled (LAS) and unlabeled (UAS) attachment scores.

Features. Given the specific experiments, we run either lexicalized or delexicalized parsers. We train lexicalized parsers using the following features, which relate to CoNLL-X specifications: word forms (FORM), coarse-grained POS tags (CPOS), morphological features (FEATS), and the dependencies (HEAD, DEPREL). In delexicalized parsing, we drop the lexical features (FORM), and the morphological features (FEATS), to arrive at the single-source delexicalized transfer parsing baseline of McDonald et al. (2013). As the focus of our assessments lies exclusively in dependency parsing, we do not experiment with POS tagging, and we use gold POS tags in all experiments, as well as gold morphological features. For a detailed account on the predicted tag impact in parsing Croatian and Serbian, see (Agić et al., 2013b), and note here that the decrease is easily quantifiable at 2-3 points LAS on average.

Data. In the first batch of experiments, we train the parsers on the 3,557 sentences from SETIMES.HR and Croatian UD, i.e., we omit the development set from all runs. In the second batch, we use the source treebanks from the CoNLL 2006-2007 datasets (Buchholz and Marsi, 2006; Nivre et al., 2007), and the UD version 1.0 release.[5] The test sets always remain the same, albeit they do appear in their lexicalized or delexicalized forms: they are the 4 x 100 Croatian and Serbian newswire (NEWS) and Wikipedia (WIKI) samples.

Next, we provide a more detailed insight into the experiments as we discuss the results of the two batches.

3.2 Croatian as Source

Here, we train parsers on Croatian training data, and evaluate them on Croatian and Serbian test sets. We parse with the SETIMES.HR data and

[4] https://code.google.com/p/mate-tools/

[5] http://hdl.handle.net/11234/1-1464

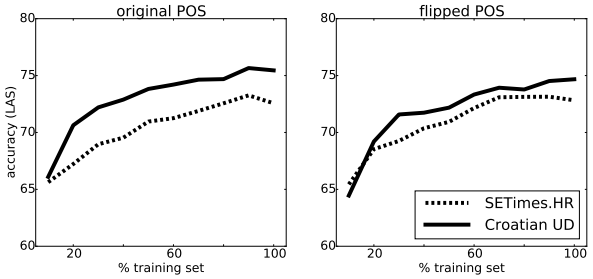

Figure 2: Learning curves (LAS) for the two treebanks with original and exchanged POS annotations. Tested on the merged test sets.

MTE4 features, as well as with the UD data and UPOS features. As for the features, we compare the POS-only setups to the setups using POS and full morphological features. The results are presented in Table 2. Note that we should not (and do not) directly compare SETIMES.HR and UD accuracies, as they are not directly comparable due to different annotation schemes.

Overall—on the merged Croatian + Serbian test sets—the parser scores at 76.4 points LAS with the best SETIMES.HR model, the one using full MTE4 morphology. Around 4 points are lost when dropping the morphology and using only POS. The system performs the best on in-domain newswire data, and records drops when moving out to Wikipedia text. Accuracies on Croatian and Serbian data are virtually identical on average, with slight preference to Croatian in-domain and Serbian out-of-domain text. Identical patterns hold for the UD experiments as well, but the scores surpass the previous ones by 2-4 points LAS, reaching the average accuracy of 79.6 points LAS for parsing Croatian and Serbian with UD. This is the highest reported score for parsing Croatian and Serbian so far, cf. Agić et al. (2014b). The average gain from adding full UD morphology on top of UPOS amounts to 4.3 points LAS. All UAS scores reported in Table 2 correspond to their respective LAS patterns.

To actually compare UD to SETIMES.HR, we perform another experiment. Since the same text is annotated twofold in our treebank—with two sets of morphological and syntactic annotation layers—we control for the morphological annotation to observe its effects on parsing. Namely, in the Table 2 report, we used each syntactic layer with its native morphological layer: SETIMES.HR with MTE4, and UD with UPOS. Now, we flip the morphology, and report the scores: we parse for

	CoNLL		UD			
	hrv	srp	hrv		srp	
Source	*UAS*	*UAS*	*UAS*	*LAS*	*UAS*	*LAS*
Bulgarian	49.8	49.2	64.1	50.6	66.6	53.8
Czech	36.3	36.1	69.9	54.8	71.9	57.3
Danish	42.1	42.2	56.7	44.2	56.9	45.6
German	40.6	41.5	58.1	41.8	60.0	45.1
Greek	61.7	63.4	52.0	32.8	53.8	35.1
English	46.3	46.5	54.6	41.3	57.1	44.1
Spanish	30.4	33.5	60.8	43.7	64.1	47.5
French	40.3	42.7	56.6	41.4	56.3	42.3
Italian	43.2	45.0	61.3	45.5	62.5	47.6
Swedish	40.2	41.2	55.9	42.7	56.4	44.4
AVERAGE	43.1	44.1	59.0	43.9	60.6	46.3

Table 3: Cross-lingual parsing accuracy for the delexicalized parsers on Croatian (hrv) and Serbian (srp) as targets. We highlight the best CoNLL and UD scores separately.

SETIMES.HR syntax by using UPOS features, and for UD syntax by using MTE4 features. This way, we get to see whether the difference in LAS scores is accounted for by the morphological features, or facilitated by the annotation schemes themselves. We report this experiment in the form of learning curves in Figure 2. We notice that SETIMES.HR parsing does not benefit at all from using the UPOS features, as the scores remain virtually identical. In contrast, the UD parsing accuracy slightly decreases when using MTE4 instead of UPOS, while still maintaining the edge over SETIMES.HR. From this we conclude that 1) the decrease in the UD scores reflects the better parsing support provided by UPOS in comparison to MTE4, and that 2) the SETIMES.HR scheme is inherently harder to parse, since it plateaus for both POS feature sets, while UD benefits from the change (back) to UPOS. The first observation is unsurprising given that UPOS differentiates, e.g., between main and auxiliary verbs, or common and proper nouns, while MTE4 POS does not. The second observation is much more interesting, especially given the syntactic tagset differences, as there are only 15 tags in SETIMES.HR, and 39 in Croatian UD. The result seems to indicate that UD outperforms SETIMES.HR without sacrificing the expressivity. However, we do note—following Elming et al. (2013)—that our evaluation is intrinsic, and that the two treebanks should be compared on downstream tasks that require parses as input.

3.3 Croatian and Serbian as Targets

In this experiment, we basically replicate the single-source delexicalized transfer setups of (McDonald et al., 2011; McDonald et al., 2013), but with Croatian and Serbian as target languages. We select ten languages with treebanks in both the CoNLL 2006-2007 datasets and the UD version 1.0 release, making for 2 x 10 = 20 different treebanks. We delexicalize the treebanks, keeping CPOS the only observable feature, and train the delexicalized parsers. Finally, we apply the parsers on the Croatian and Serbian test sets, evaluating for attachment scores.

Before discussing the scores, we record a few relevant details about our setup. First, we only parse the SETIMES.HR test sets using the CoNLL models, and the UD test sets using the UD models. This is to illustrate the difference between evaluating cross-lingual parsers in heterogenous and homogenous environments regarding the treebank annotations, but now with an outlook on Croatian and Serbian. Second, building on that setup, we only evaluate the CoNLL parsers for UAS, while the UD parsers are inspected for both UAS and LAS, as the syntactic tagsets do not overlap between the CoNLL datasets or with the SETIMES.HR tagset. In contrast, the core UD tag collection is uniform across the languages. Third, the CoNLL datasets we use are the POS tags of Petrov et al. (2012), so we map the UPOS tags to those in all our CoNLL experiments. The mapping itself is trivial, as UPOS is a simple extension of the (Petrov et al., 2012) tagset. Fourth and final, all ten source languages are European by virtue of overlapping CoNLL and UD, and not by deliberately excluding other datasets. The group does have typological subsets of interest for cross-lingual parsing of Croatian and Serbian.

Our observations for transferring the CoNLL parsers are consistent with those of McDonald et al. (2011): the accuracies do not seem to bear any typological significance, and the scores are relatively low, signalling underestimation. The best cross-lingual parser seems to be the one induced from the Greek treebank, while those of more closely related Slavic languages—Bulgarian and Czech—fall far behind in scores. Actually, in this scenario, Czech is the second worst choice for parsing Croatian and Serbian, in spite of having a very large and consistently annotated treebank. This is apparently due to the treebank heterogeneity, as we know from a large body of related work from McDonald et al. (2011) on.

In contrast to the CoNLL scores, the UD parsers perform much better, and in much more accordance with our typological intuitions. The best two parsers are trained on Bulgarian and Czech data, the latter one scoring a notable 69.9 and 71.9 points UAS on Croatian and Serbian. The LAS scores are expectedly much lower, and the accuracies are consistent with related work (McDonald et al., 2013; Agić et al., 2014b). On average, the UD treebanks score 15 or more points UAS above the CoNLL treebanks. This figure in itself only instantiates the concerns with evaluating parsers on heterogenous resources, and the alleviation of these concerns via resource uniformity. On top of that, we establish a typological ordering of ten languages as sources in parsing Croatian and Serbian.

4 Related Work

Tadić (2007) marks the beginning of Croatian treebanking by discussing the applicability of the Prague Dependency Treebank (PDT) syntactic annotation scheme (Böhmová et al., 2003) for Croatian, supporting the discussion with a small sample of 50 manually annotated Croatian sentences dubbed the Croatian Dependency Treebank (HOBS). By the time parsing experiments of Berović et al. (2012) and Agić (2012) were conducted, HOBS already consisted of more than 3,000 sentences. Its latest instance—complete with Croatian-specific annotations of subordinate clauses, but otherwise fully PDT-compliant—encompasses 4,626 sentences of Croatian newspaper text (Agić et al., 2014a). A version of HOBS is available under a non-commercial license.[6]

SETIMES.HR is a treebank of Croatian built on top of the newspaper text stemming from the SETIMES parallel corpus of SEE languages.[7] It was built to facilitate accurate parsing of Croatian through a simple dependency scheme, and also to encourage further development of Croatian resources via very permissive free-culture licensing. The treebank currently contains approximately 9,000 sentences, and it is freely available for all purposes. Agić and Ljubešić (2014) observe state-of-the-art scores in Croatian lemmatization, tagging, named entity classification, and dependency parsing using SETIMES.HR with stan-

[6] http://meta-share.ffzg.hr/

[7] http://opus.lingfil.uu.se/SETIMES.php

dard tools. Furthermore, this line of research explores the usage of Croatian resources as sources for processing Serbian text (Agić et al., 2013a; Agić et al., 2013b), and also the possibility of sharing models between SEE languages (Agić et al., 2014b). These experiments result in promising findings regarding model transfer between related languages, and they bring forth state-of-the-art scores in processing Croatian, Serbian, and Slovene, offerring freely available resources.

Given the extensive lines of work in Croatian treebanking—with three different reasonably-sized dependency treebanks, cross-domain test sets, and practicable accuracies—it is safe to argue that Croatian is departing the company of severely under-resourced languages when it comes to dependency parsing. In contrast, Serbian treebanking is at this point virtually non-existent. To the best of our knowledge, its only reference point seems to be a study in preparing the morphological annotations for a future—possibly also PDT-compliant—dependency treebank of Serbian (Djordjević, 2014). In absence of such a treebank, Agić et al. (2014b) provide state-of-the-art scores in Serbian parsing using the PDT and SETIMES.HR schemes, while our work presented in this paper offers a very competitive UD parser for Serbian via direct transfer from Croatian.

5 Conclusions

We have presented a new linguistic resource for Croatian: a syntactic dependency treebank within the Universal Dependencies framework. It consists of approximately four thousand sentences, and comes bundled with two-domain test sets for Croatian and Serbian. It is built on top of an existing treebank of Croatian, the SETIMES.HR corpus. We have intrinsically evaluated the resources in a monolingual parsing scenario, as well as through cross-lingual delexicalized transfer parsing into Croatian and Serbian using twenty different source parsers. We recorded state-of-the-art performance in parsing the two languages, at approximately 80 points LAS. All the resources used in the experiment are made publicly available: `https://github.com/ffnlp/sethr`.

Future work. We have described the first instance of Croatian UD. We seek to improve the resource in many ways, and to utilize it in experiments featuring dependency parsing. The treebank is currently not documented, and we aim at providing proper documentation via the UD platform for the next release. Moreover, we currently do not make use of any language-specific features in morphology and syntax. Following the experience of other Slavic languages in the UD project, we might augment the Croatian annotations with language specifics as well. Finally, albeit not exclusively, the research in Croatian parsing and sharing resources between the SEE languages requires extensive downstream evaluation, which we hope to provide in future experiments, together with resources facilitating future downstream evaluations for these languages.

Acknowledgements. We thank the anonymous reviewers for their comments. We also acknowledge the efforts of our annotators in producing the first version of the syntactic annotations, and in facilitating the process of UD adoption for Croatian.

References

Anne Abeillé. 2003. *Treebanks: Building and Using Parsed Corpora*. Springer.

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *LREC*, pages 1724–1727.

Željko Agić and Danijela Merkler. 2013. Three syntactic formalisms for data-driven dependency parsing of Croatian. *LNCS*, 8082:560–567.

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013a. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *BSNLP*, pages 48–57.

Željko Agić, Danijela Merkler, and Daša Berović. 2013b. Parsing Croatian and Serbian by using Croatian dependency treebanks. In *SPMRL*, pages 22–33.

Željko Agić, Daša Berović, Danijela Merkler, and Marko Tadić. 2014a. Croatian dependency treebank 2.0: New annotation guidelines for improved parsing. In *LREC*, pages 2313–2319.

Željko Agić, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkler, and Sara Može. 2014b. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *LT4CloseLang*, pages 13–24.

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci,

Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.1.

Željko Agić. 2012. K-best spanning tree dependency parsing with verb valency lexicon reranking. In *COLING*, pages 1–12.

Daša Berović, Željko Agić, and Marko Tadić. 2012. Croatian dependency treebank: Recent developments and initial experiments. In *LREC*, pages 1902–1906.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Springer.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *COLING*, pages 89–97.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *CoNLL*, pages 149–164.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*, pages 600–609.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, pages 4585–4592.

Bojana Djordjević. 2014. Initial steps in building Serbian treebank: Morphological annotation. In *Natural Language Processing for Serbian: Resources and Applications*, pages 41–53.

Jakob Elming, Anders Johannsen, Sigrid Klerke, Emanuele Lapponi, Hector Martinez Alonso, and Anders Søgaard. 2013. Downstream effects of tree-to-dependency conversions. In *NAACL*, pages 617–626.

Tomaž Erjavec. 2012. Multext-East: Morphosyntactic resources for central and eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

Richard Johansson. 2013. Training parsers on incompatible treebanks. In *NAACL*, pages 127–137.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *ACL*, pages 92–97.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *EMNLP-CoNLL*, pages 915–932.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.0.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC*, pages 2089–2096.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *SPMRL*, pages 146–182.

Marko Tadić. 2007. Building the Croatian dependency treebank: The initial stages. *Suvremena lingvistika*, 63:85–92.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *COLING*, pages 1854–1864.

Hans Uszkoreit and Georg Rehm. 2012. *Language White Paper Series*. Springer.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT*, pages 1–8.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtskỳ, and Jan Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.