

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

Željko Agić* **Nikola Ljubešić**†

* Center for Language Technology
University of Copenhagen, Denmark

† Dept. of Information and Communication Sciences,
Faculty of Humanities and Social Sciences,
University of Zagreb

BSNLP 2015, 10th Sep 2015

Introduction

- for parsing we need supervision in form of annotated corpora
- dependency treebanks costly to develop and follow different annotation schemes across languages
- this hinders cross-lingual parsing and enabling LT for under-resourced languages
- Universal Dependencies [Nivre et al., 2015] address this issue by providing homogenous dependency treebanks
- parts of speech, morphological features and syntactic annotations across 18 languages
- [McDonald et al., 2013] stress the two obvious gains from uniform schemata:
 - ① more exact evaluation of dependency parsers
 - ② typologically motivated transfer of dependency parsers to under-resourced languages

Contributions

- focus on cross-lingual dependency parsing of two under-resourced South Slavic languages
- ① dependency treebank for Croatian
- ② cross-domain test sets for Croatian and Serbian
- ③ set of experiments for parsing the languages within the UD framework
- ④ cross-lingual parsing experiments, target Croatian and Serbian by source models from 10 treebanks, two types (CoNLL and UD)
- ⑤ make our datasets available under free-culture licensing
<https://github.com/ffnlp/sethr>

The treebank

- built on top of the SETIMES.HR dependency treebank [Agić and Ljubešić, 2014]
- 3,557 training sentences (newswire)
- 200 dev sentences from same source
- 400 test sentences
 - 200 Croatian, 200 Serbian
 - 200 from same source, 200 from Wikipedia
 - 100 per source and language
- implement the following annotation layers (first two mandatory):
 - ① universal POS tags
 - ② dependency attachment
 - ③ universal morphological features

Morphology

- SETIMES.HR implements (a revision of) the Multext East version 4 morphosyntactic tagset (MTE4) [Erjavec, 2012]
- manually convert it to
 - UD's universal POS tags (UPOS)
 - universal morphological features
- out of 17 UPOS tags 14 used in our treebank
- leave out determiners (DET), interjections (INTJ), and symbols (SYM)
- MTE4 abbreviations mapped context-dependent to appropriate UPOS tags, mostly nouns, but adverbs as well (“npr.” = “e.g.”)
- conflate the 1316 seen tags to 14

Experimental setup

- two sets of experiments
 - ① Croatian as source – monolingual parsing of Croatian and transfer to Serbian
 - ② Croatian and Serbian as target – transfer of delexicalised parsers from 10 well-resourced languages to Croatian and Serbian
- parser – *mate-tools* graph-based parser of [Bohnet, 2010]
- evaluation – LAS and UAS
- features
 - word form (FORM)
 - coarse-grained POS tag (CPOS)
 - morphological features (FEATS)
 - dependencies (HEAD, DEPREL)
- delexicalised parser drops FORM and FEATS

Croatian as source

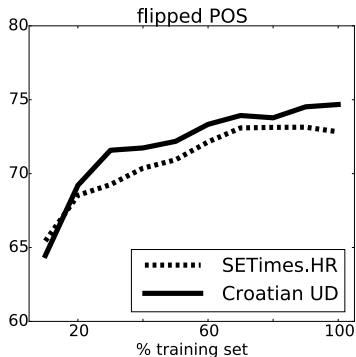
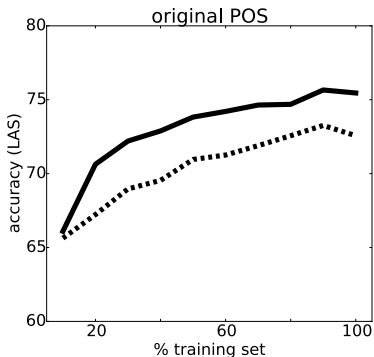
- train on the Croatian train set, evaluate on Croatian and Serbian test sets

Treebank	Features	Croatian				Serbian			
		NEWS		WIKI		NEWS		WIKI	
		<i>UAS</i>	<i>LAS</i>	<i>UAS</i>	<i>LAS</i>	<i>UAS</i>	<i>LAS</i>	<i>UAS</i>	<i>LAS</i>
SET.HR	CPOS	82.2	76.3	77.1	67.9	80.8	74.0	79.8	71.1
	+ FEATS	84.3	79.2	80.7	73.7	83.0	77.8	82.6	74.7
UD	CPOS	84.8	77.9	80.8	72.4	82.4	75.8	82.1	75.2
	+ FEATS	86.9	81.5	84.5	77.3	86.0	81.5	83.7	77.9

- morphological features add consistently 2-4 points
- UD outperforms SETIMES.HR for 2-3 points?

UD vs. SETIMES.HR

- flip POS information to observe the impact of the syntactic layer only



- for any final conclusions the parser outputs still have to be evaluated extrinsically on downstream tasks!

Croatian and Serbian as targets

- replicate the single-source delexicalised transfer setups of [McDonald et al., 2011, McDonald et al., 2013] – CPOS the only observable feature
- select 10 languages with treebanks in both CoNLL 2006-2007 and UD v1.0
- evaluate CoNLL on `SETIMES.HR` – heterogenous setting
- UD evaluated on UD – homogenous
- evaluate CoNLL on UAS only as CoNLL and `SETIMES.HR` labels do not overlap
- for CoNLL experiments map the UPOS to [Petrov et al., 2012]

Croatian and Serbian as targets

Source	CoNLL		UD			
	hrv	srp	hrv		srp	
	<i>UAS</i>	<i>UAS</i>	<i>UAS</i>	<i>LAS</i>	<i>UAS</i>	<i>LAS</i>
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

Conclusion and future work

- presented the Croatian syntactic dependency treebank within the Universal Dependencies framework
- cca. 4,000 sentences with two-domain two-languages test sets
- intrinsic evaluation via monolingual parsing with ~ 80 LAS on both languages
- although the label set is twice the size, UD proven to be easier to parse than SETIMES.HR
- heterogenous vs. homogenous delexicalised cross-lingual parsing – homogenous gives much better results, following typological similarities
- future work
 - writing UD documentation
 - currently do not utilise language-specific features in neither morphology nor syntax
 - downstream evaluation!

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

Željko Agić* **Nikola Ljubešić**†

* Center for Language Technology
University of Copenhagen, Denmark

† Dept. of Information and Communication Sciences,
Faculty of Humanities and Social Sciences,
University of Zagreb

BSNLP 2015, 10th Sep 2015