Parsing Croatian and Serbian by Using Croatian Dependency Treebanks

Željko Agić* **Danijela Merkler†** **Daša Berović†**
*Department of Information and Communication Sciences
†Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
zagic@ffzg.hr dmerkler@ffzg.hr dberovic@ffzg.hr

Abstract

We investigate statistical dependency parsing of two closely related languages, Croatian and Serbian. As these two morphologically complex languages of relaxed word order are generally under-resourced – with the topic of dependency parsing still largely unaddressed, especially for Serbian – we make use of the two available dependency treebanks of Croatian to produce state-of-the-art parsing models for both languages. We observe parsing accuracy on four test sets from two domains. We give insight into overall parser performance for Croatian and Serbian, impact of preprocessing for lemmas and morphosyntactic tags and influence of selected morphosyntactic features on parsing accuracy.

1 Introduction

Croatian and Serbian are very closely related South Slavic languages with complex morphology and relatively free word order. They are mutually intelligible with one another, as well as with Bosnian and Montenegrin, amounting for more than 20 million native speakers.[1] Regarding language technology support, they are considered to be generally under-resourced. More specifically, while a corpus of research on processing Croatian and Serbian on the morphosyntactic and shallow syntactic layer does exist (Tadić et al., 2012; Vitas et al., 2012), approaches to full syntactic analysis of the two languages were up to this point very sparse and very recent (Agić and Merkler, 2013). As linguistic tradition supports dependency-based syntactic formalisms for the two languages (Böhmová et al., 2003; Tadić, 2007), it should be noted that they have not participated in the previous collaborative research efforts in dependency parsing, such as the CoNLL shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007). Furthermore, regardless of the specific research topic, the communities dealing with natural language processing of Croatian, Serbian and other closely related languages from their group are still to reach the common level of awareness with respect to public availability of their research. Contributions to availability of Croatian and Serbian resources have once again been very few and recent (Tadić and Varadi, 2012), especially for free culture licensing.

Through the line of research we propose here,[2] we seek to provide state-of-the-art in dependency parsing for both Croatian and Serbian. In this first group of experiments, we build on the fact of their close relatedness by using the two Croatian treebanks – Croatian Dependency Treebank (Tadić, 2007) and SETIMES.HR Treebank (Agić and Merkler, 2013) – to build unified parsing models and evaluate them across the languages and domains. As we deal with highly inflectional languages, we also investigate the influence of morphological preprocessing and morphosyntactic feature selection on parsing perfor-

[1] Bekavac et al. (2008) provide a corpus-based comparison of Bosnian, Croatian and Serbian, observing similarities and differences in morphology, syntax and semantics. For further insight regarding Croatian and Serbian morphosyntax, see the respective contemporary grammars (Silić and Pranjković, 2005; Stanojčić and Popović, 2008).

[2] This work was partly financed by the EU FP7 STREP project XLike (FP7-288342).

Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages, pages 22–33,
Seattle, Washington, USA, 18 October 2013.

mance. We aim to use this first inquiry as a decision point regarding further advancements in resource interchangeability in terms of, e.g., annotation projection (Yarowsky et al., 2001) and domain adaptation (Søgaard, 2013). Availability is highly emphasized, as we provide our resources and models to the public under the CC-BY-SA-3.0 license.[3] We stress the essential role of free culture licensing in enabling and maturing NLP for under-resourced languages.

In the following section, we give an overview of related work in computational processing of Croatian and Serbian morphology and syntax. Further, we define the experiment objectives and describe the resources and experiment workflow. We elaborate on the obtained results and conclude by sketching possible future research plans.

2 Related work

Two overviews of current state of language technology development have appeared just recently for the two languages we investigate in this paper.

The Croatian overview (Tadić et al., 2012) states that a few underperforming shallow parsing prototypes for Croatian do exist (Vučković et al., 2008), while deep parsing is left completely unaddressed. In contrast, it indicates that the more basic resources – manually annotated corpora, inflectional lexicons, lemmatizers, morphosyntactic and named entity taggers – are of higher quality and availability. Most of these are available through META-SHARE (Tadić and Varadi, 2012). However, in terms of mandatory preprocessing for dependency parsing, to the best of our knowledge, the only freely available and standard-compliant lemmatization, part-of-speech (POS) or morphosyntactic (MSD) tagging resources are those by (Agić et al., 2013).[4] Their elaboration contains a more substantial overview of preprocessing. Relevant to our research, these models provide the state of the art in preprocessing for both Croatian and Serbian.

Croatian Dependency Treebank (HOBS) project was initiated by (Tadić, 2007). However, its sufficiency in size increase, followed by the first experiments with dependency parsing of Croatian, did not appear soon enough to be included in the CoNLL shared tasks and the overview of (Tadić et al., 2012). Preliminary experiments in transition-based (Berović et al., 2012) and graph-based parsing have been augmented by a hybrid approach which included integrating a graph-based parser (Hall, 2007) and a valency lexicon (Agić, 2012). Due to uncovered partial inadequacies of the HOBS formalism at describing certain syntactic properties of Croatian, a new line of research was initiated, aiming at creating a more simplistic dependency-based formalism for data-driven parsing of Croatian (Agić and Merkler, 2013). It provided a new freely available dependency treebank, the SETIMES.HR Treebank, and derived state-of-the-art dependency parsing models.[5] On the downside, SETIMES.HR is a prototype with currently less than 2 500 sentences and a documented need for addressing certain annotation challenges, such as consistent annotation of complex predicates, an issue that was previously observed and partially resolved in HOBS as well (Berović et al., 2012).

The overview of Serbian language technologies (Vitas et al., 2012) explicitly denotes a satisfactory development level for Serbian preprocessing based on large electronic dictionaries, manually annotated corpora and hand-crafted transducer grammars. These are available through META-SHARE, even if mostly coupled with restrictive licensing. Further, the overview lists some preliminary research in shallow syntactic analysis, while it clearly states that the absence of a formalised syntax of Serbian restricts the development of syntactically annotated corpora and thus hinders the research in full parsing of Serbian, making the creation of a syntactic formalism for Serbian a very urgent task.

Similar to Croatian, research in Serbian shallow parsing deals exclusively with the manual design of rule-based modules (Nenadić, 2000; Nenadić et al., 2003; Vitas et al., 2003) in linguistic development enviroments such as Intex and NooJ (Silberztein, 2004). We also inquired into a case study on the possibilities of resource transfer from English to Serbian (Martinović, 2008), only to conclude that it does not provide any empirical results. Hence, to the best of our knowledge, no experiments in dependency treebank construction and data-driven de-

[3] http://creativecommons.org/licenses/by-sa/3.0/

[4] http://nlp.ffzg.hr/resources/models/tagging/

[5] http://nlp.ffzg.hr/resources/corpora/setimes-hr/

pendency parsing – or, for that matter, any other approaches to deep syntactic modeling and processing – currently exist for Serbian.

3 Experiment setup

In this section, we present the experimental setup by which we aim at subsequently addressing the previously outlined issues with dependency parsing of Croatian and Serbian. We define our goals, describe the utilized resources and lay out the workflow.

3.1 Objectives

We identify the main issues unaddressed by previous research in Croatian and Serbian syntactic processing and use these to define our research objectives. They are listed here as follows.

1. No empirical research was conducted in dependency parsing of Serbian. Even if this fact was justified by the lack of applied research in creating formalisms targeted exclusively at describing syntactic properties of Serbian, we follow the underspecification approach that was successfully implemented in HOBS for Croatian. Namely, as the Prague Dependency Treebank (PDT) formalism for Czech (Böhmová et al., 2003) was altogether ported to Croatian by simply using the PDT annotation manual for annotating Croatian sentences due to minor differences in syntactic structure between Croatian and Czech, we reflect this to the even greater similarity between Croatian and Serbian on all levels of linguistic description. Hence, we use Croatian data to parse Serbian and to serve as a baseline in Serbian parsing.
2. Using Croatian syntactic models for parsing Serbian text serves to establish the need for advanced approaches to porting resources among languages, such as annotation projection.
3. The best dependency parsing models for Croatian are created and tested using a small prototype treebank. SETIMES.HR currently provides state of the art in Croatian dependency parsing. To serve our experiment, we enlarge it by 50% by following the annotation guidelines (Merkler et al., 2013) and provide its new version to the public.
4. Previous experiments were conducted by tenfold cross-validation on treebank data. This is a standard approach to dependency parser evaluation, especially in under-resourced environments. In this setting, observations are positively biased by text domain and phrase transfer due to randomization. We seek to partially account for these effects by designing a set of language- and domain-aware test samples. By these we also target at establishing the need for domain adaptation for parsing.
5. No research was done in investigating the effects of preprocessing and linguistic feature selection to dependency parsing for these languages. As these are highly inflectional, having very large morphosyntactic tagsets, we seek to inspect the impact of preprocessing choices on their dependency parsing. There is ample research on the effect preprocessing has on dependency parsing (Goldberg and Elhadad, 2009; Mohamed, 2011) and on joint morphological and syntactic processing (Bohnet and Nivre, 2012), but none of it included any of the South Slavic languages.

3.2 Workflow

We define three batches of experiments to meet the research objectives:

1. to select the best Croatian dependency formalism with respect to its overall parsing accuracy on Croatian and Serbian – with an emphasis on the most important syntactic categories that match across formalisms – and incidentally to establish the need for annotation projection,
2. to inspect the impact of state-of-the-art automatic preprocessing on dependency parsing of both languages and
3. to establish the importance of specific Croatian and Serbian morphosyntactic features of the most frequent parts of speech in modeling syntactic fenomena for dependency parsing.

In the first batch, we use HOBS in two instances and SETIMES.HR to create parsing models and test them on Croatian and Serbian test samples. Drawing from previous research, we use a standard non-projective graph-based MSTParser generator with second-order features (McDonald et al., 2006), as this setting favors Croatian (Agić, 2012) and re-

lated languages such as Czech and Slovene (Buchholz and Marsi, 2006). We are aware of the existence of novel dependency parsers that implement approaches to handling non-local dependencies and outperform MSTParser on a set of languages, such as (Bohnet and Nivre, 2012). They are not included here due to temporal constraints and the fact that we were provided with prebuilt MSTParser models for the HOBS instances and needed to ensure their comparability with SETIMES.HR. As we mainly deal with the concept of resource sharing between closely related languages, we assign a more elaborated parser selection for future research.

For the second batch, we redo the experiments from the first batch in a realistic scenario regarding preprocessing. We use the publicly available state-of-the-art tagging and lemmatization models for Croatian and Serbian (Agić et al., 2013) instead of manual annotation to observe the incurred effects. We do both batches for all three formalisms (two HOBS instances and SETIMES.HR) and provide learning curves.

The third batch of experiments deals with observing the impact of certain morphosyntactic features by removing them from training and test data. We inspect all features involved in subspecification of adjectives, nouns and verbs in compliance with the Multext East specification (Erjavec, 2012), i.e., MTE v5 as its fifth release.[6]

In all batches, we observe labeled (LAS) and unlabeled (UAS) attachment scores. We use approximate randomization for statistical significance testing where applicable and meaningful.

3.3 Treebanks

Two Croatian dependency treebanks are used in this experiment: HOBS (Tadić, 2007) and SETIMES.HR (Agić and Merkler, 2013).

HOBS is available in two instances or implementations. The first one closely follows the PDT annotation guidelines (Böhmová et al., 2003) with several adaptations of predicate annotation (Berović et al., 2012). The second one introduces a set of additional syntactic tags used for the introduction and subclassification of subordinate clauses. It also alters the head attachment rules for subordinating conjunctions. This addition enabled consistency in predicate annotation in clauses and an increase in dependency parsing accuracy (Agić and Merkler, 2013), while taking a turn away from the PDT guidelines and towards specifics of Croatian syntax. In the paper, we refer to this instance of HOBS as HOBS + Sub. Both of them are based on Croatian newspaper text and manually preprocessed. They implement a morphosyntactic tagset based on, but slightly deviated from MTE v4 (Erjavec, 2012). HOBS is available from META-SHARE for research purposes, but its syntactic tags are stripped from this version. HOBS + Sub is not publicly available. Both have been made available to us in whole for conducting this experiment, along with prebuilt MSTParser models compatible with our experimental settings.

SETIMES.HR is based on Croatian newspaper text

[6] http://nl.ijs.si/ME/V5/msd/html/

Features	HOBS	HOBS + Sub	SETIMES.HR
Sentences	4 626	4 626	3 853
Tokens	117 369	117 369	86 991
Types	25 038	25 038	17 723
Lemmas	12 388	12 388	8 773
MSD tags	914	911	662
Syn. tags	27 (70)	28 (81)	15

Table 1: Basic treebank statistics. Syntactic tag counts are given for the basic and the full tagset (the latter inside brackets) for the two HOBS treebanks.

	set.test		wiki.test	
Features	hr	sr	hr	sr
Sentences	100	100	100	100
Tokens	2 285	2 308	1 878	1 947
Types	1 265	1 246	1 027	1 055
Lemmas	989	979	803	797
MSD tags				
MTE v4 tags	236	237	189	193
MTE v5 tags	233	234	192	195
Syntactic tags				
HOBS	22(37)	23(37)	22(41)	22(44)
HOBS + Sub	22(46)	24(49)	23(49)	22(50)
SETIMES.HR	15	15	15	15

Table 2: Basic statistics for the four test sets. Morphosyntactic and syntactic tag counts are given with respect to the formalism used.

from the SETimes parallel corpus.[7] It implements a simplistic new formalism (Merkler et al., 2013) targeting and reaching increased dependency parsing performance while maintaining the information on the main syntactic categories and compliance with the general guidelines for HOBS for these categories (Agić and Merkler, 2013). It is also manually preprocessed, but using the newer MTE v5 morphosyntactic tagset. SETIMES.HR is fully compliant with this tagset. As mentioned, it is freely available for all purposes. With this in mind, following the annotation guidelines, we have expanded its 2 500 sentence prototype by introducing 1 365 new sentences.

Treebank statistics are given in Table 1. HOBS treebanks are larger than SETIMES.HR by approximately 800 sentences, i.e., 30 thousand tokens (30 kw). The morphosyntactic tagsets also differ, favoring SETIMES.HR and MTE v5 by 250 tags if we are to consider the smaller tagset as better in terms of the expressivity vs. preprocessing accuracy balancing. Syntactic tagset of SETIMES.HR has only 15 tags. Tag counts for HOBS treebanks are given by two figures: the first one represents the basic tagset, while the second one includes the subclassification tags. For example, a coordinated predicate is annotated as *Pred* using the basic tagset and as *Pred_Co* in the full tagset. Here, we use only the basic tagset.

As we anticipated given the properties of Croatian syntax, non-projectivity is amply present in both treebanks. Approximately 2% of all dependency relations and more than 20% of all sentences are non-projective, supporting our parser selection.

As the three treebanks – HOBS, HOBS + Sub and SETIMES.HR– formally do implement different approaches to syntactic modeling, issues may be raised regarding the comparability of dependency parsing scores. However, since HOBS and HOBS + Sub are both based on the PDT formalism and SETIMES.HR implements a simplistic formalism that is still based on the PDT and HOBS annotation guidelines and syntactic tagset reduction (Merkler et al., 2013), we consider the comparison to be valid. Moreover, all three formalisms encode the main Croatian syntactic categories by closely following the general guidelines for describing the Croatian syntax (Silić and Pranjković, 2005), thus indicating that comparisons for the main syntactic categories – such as predicates, subjects, objects, prepositional and adverbial phrases – should hold true for the task of dependency parsing irregardless of the formal differences between the models.

3.4 Test sets

The publicly available test sets are obtained from an experiment in lemmatization and tagging of Croatian and Serbian (Agić et al., 2013). They were available in MTE v4 and v5. As HOBS uses the former and SETIMES.HR the latter tagset, they were well-suited for our experiment. We syntactically annotated the test sets threefold, i.e., by using the HOBS, HOBS + Sub and SETIMES.HR formalisms. There are four test samples: Croatian and Serbian parallel sentences from newspaper sources (set.test) and Wikipedia (wiki.test). Their suitability for testing models on closely related languages was thoroughly elaborated by (Agić et al., 2013), where their differences were measured by using inflectional lexicons of Croatian and Serbian and were found to be significant in supporting the difference between the languages. Namely, lexical coverage differed by approximately 10 percentage points in favor of Croatian across the two domains.

Statistics for the test set are given in Table 2. Each sample has 100 sentences or approximately 2 000 tokens. Slight variations in token, type and lemma counts are present and reflect the domain differences. MSD tag and syntactic tag counts reflect the respective formalisms, as not all HOBS and HOBS + Sub syntactic tags are utilized, while all 15 SETIMES.HR tags are present in all the samples. HOBS tag counts are once again given separately for the basic and the full tagset, while only the basic subset was used in the experiment.

Inter-annotator agreement for HOBS, HOBS + Sub and SETIMES.HR is investigated in (Agić and Merkler, 2013). It favors SETIMES.HR over HOBS + Sub and HOBS + Sub over HOBS with a statistically significant difference. The CoNLL shared tasks in dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007) used test sets of approximately 5 000 tokens. This may raise an issue regarding the relatively small size of our domain test samples. However, in the experiment, we combine the test sets by domain and by language and also

[7] http://opus.lingfil.uu.se/SETIMES2.php

	set.test		wiki.test		
LAS	hr	sr	hr	sr	overall
HOBS	59.9	58.7	55.5	55.4	57.6
HOBS + Sub	68.3	66.9	62.4	62.7	65.3
SETIMES.HR	**76.7**	**75.4**	**71.9**	**72.4**	**74.3**
UAS					
HOBS	73.7	75.9	72.3	72.6	73.8
HOBS + Sub	78.1	79.0	76.5	76.5	77.6
SETIMES.HR	**81.6**	**80.6**	**80.0**	**80.6**	**80.8**

Table 3: Parsing accuracy (LAS, UAS) with manual preprocessing. Results are given for each test set and overall, i.e., with all four test sets merged into one.

merge them into a single test set, thus accounting for the size of the individual samples.

3.5 Parser setup

Here we use MSTParser with the non-projective maximum spanning tree parsing algorithm and second order features (`decode-type:non-proj order:2 training-k:5 iters:10`), as it was previously established as the optimal setting for parsing Croatian using MSTParser (Agić, 2012) with a statistically significant margin over the transition-based approach. In training and testing, we separate the MTE v5 MSD tags into POS (CPOSTAG) and full MSD (POSTAG). We do not separate the MSD tags into atomic features, i.e., we do not utilize the FEATS column of the CoNLL-X format. Thus the MSD tags themselves are considered as atomic features in the experiment, both for the full MTE v5 tagset and its reductions.

4 Results and discussion

Here we report and discuss the obtained results. We discuss the results in batches, as in the experiment workflow description. In addition, we give a brief linguistic analysis of the parsing errors considering the difference between the two languages and the fact that Croatian models were used for parsing both Croatian and Serbian text.

4.1 Formalism selection

In the first experiment batch, we trained the parsing models using three treebanks, HOBS, HOBS + Sub and SETIMES.HR, and tested them on our Croatian and Serbian test sets from Wikipedia and newspaper text. We present the overall scores in Table 3, the learning curves are plotted in the first diagram of Figure 1 and the accuracy for selected syntactic categories are given in Table 4.

Regarding the formalism selection process, inspecting the overall observed LAS and UAS, it is evident that models based on SETIMES.HR outperform HOBS-based models by a large margin. They outperform HOBS + Sub by approximately 9 LAS and 3 UAS points, while their overall advantage is even more substantial in comparison with the scores of basic HOBS models – approximately 17 LAS and 7 UAS points. Benefits of explicit annotation of predicates by introducing tags for subordinating syntactic conjunctions are also evident as HOBS + Sub parsers outperform HOBS by 8 LAS and 4 UAS points. These observations maintain the conclusions about the three formalisms given in previous research (Agić and Merkler, 2013).[8] Moreover, the introduction of a held-out test set further steepens these differences, as the previous tests were performed by tenfold cross-validation using treebank data only. The observed differences in overall LAS and UAS scores are shown to be significant by the approximate randomization test ($p < 0.01$).[9]

As stated in the presentation of treebanks in the previous section, since the three formalisms are closely related to one another and to the general guidelines for describing the properties of Croatian dependency syntax, we find this comparison to hold true regardless of the formal differences between the models. Moreover, since the accuracy for the PDT-based formalisms in this and previous experiments with Croatian dependency parsing (Agić and Merkler, 2013) is below the margins set by similar languages such as Czech and Slovene (Buchholz and

[8] Importance of standard compliance should be noted regarding the morphosyntactic tagset impact on the observed results. Namely, HOBS "slightly deviates" from MTE v4 by design, while still claiming *de facto* compliance. As the test sets fully comply with MTE v4 and v5, this has an effect on parsing.

[9] We test by randomly ($prob = 0.5$) inserting alternate syntactic annotations for entire test set sentences and evaluating with respect to annotation style, i.e., selecting to match the sentence annotations against HOBS, HOBS + Sub or SETIMES.HR layer in the gold standard annotation.

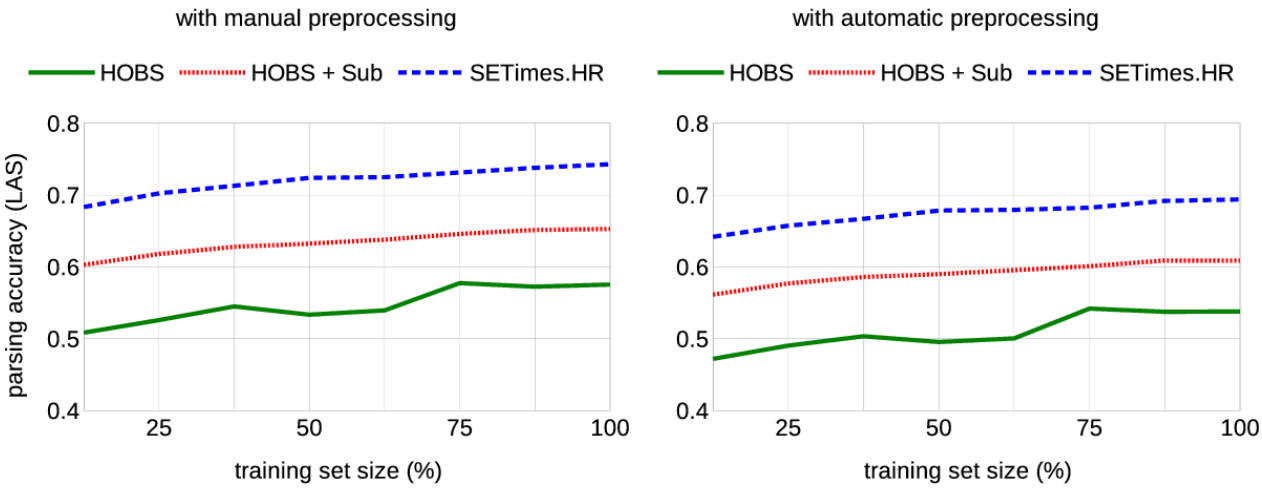

Figure 1: Labeled attachment learning curves for the three treebanks using gold standard and automatic lemmatization and morphosyntactic tagging

Marsi, 2006)[10], we argue that HOBS requires thorough further revision if it is to be the Croatian counterpart of PDT in terms of expressivity and usability in research and practical applications. This is further supported by the data in Table 4, where the assignment of specific syntactic tags is explored. However, extrinsic evaluation would also be beneficial.

The differences in LAS and UAS scores between the two languages are virtually non-existent across formalisms and domains. The parsing models favor Croatian newspaper text by less than 2 LAS points for all three formalisms, while UAS is approximately 1 UAS point higher in Serbian newspaper text for HOBS and HOBS + Sub, in contrast with SETIMES.HR, which scores 1 UAS point higher for the Croatian sample. In the Wikipedia samples, LAS and UAS may be approximated as identical. In total, as a top-performer, the SETIMES.HR model scored 74.5 LAS and 80.9 UAS on Croatian samples and 74.1 LAS and 80.6 UAS on Serbian samples. We believe this indicates that the parsing models trained on Croatian treebank data can be used reliably for both Croatian and Serbian text. We also use these figures to imply no need for syntactic annotation projection between Croatian and Serbian in this test scenario.

The cross-domain differences in LAS and UAS are, in contrast with the cross-language differences, much more substantial. As all treebanks were built on top of Croatian newspaper text, scores are expectedly higher for these test samples in comparison with the Wikipedia samples' scores. This difference amounts to approximately 5 LAS points and 2 UAS points in favor of the newspaper text samples across the two languages and three formalisms.

We plotted the LAS learning curves by merging the test samples into a single mixed-language test set, incrementally creating 8 parsing models per formalism (12.5% to 100% of full size) and testing them on this merged test set. The left plot of Figure 1 represents the learning curves for the three treebanks peaking at previously discussed scores from Table 3. The curves clearly reflect the overall differences in scores. Their rate of increase is consistently comparable, with the overall difference in favor of SETIMES.HR due to its smaller yet still informative syntactic tagset and its formalism better suited for Croatian syntax. With this fact now once again empirically supported, we select the top-performing SETIMES.HR parsing model for further inspection. Thus, our further discussion deals exclusively with parsing using SETIMES.HR.

First we observe parsing accuracy regarding syntactic categories, where we still do compare SETIMES.HR with HOBS + Sub as a final reference point. We merged our test sets by language to provide Croatian and Serbian cross-domain test samples and calculate the LAS per syntactic category for

[10]This holds even with the Slovene treebank of the CoNLL 2006 shared task having more than 2 000 sentences less than HOBS, with both using the PDT formalism

	HOBS + Sub		SETIMES.HR	
Syntactic tag	hr	sr	hr	sr
Adverb	**50.4**	46.6	**50.4**	**47.2**
Attribute	81.4	82.3	**87.9**	**88.4**
Object	56.4	51.3	**68.9**	**70.2**
Predicate	75.1	71.9	**80.7**	**81.2**
Preposition	65.5	**66.4**	**66.4**	64.0
Subject	70.3	71.3	**74.8**	**77.6**

Table 4: LAS for main syntactic tags separated for Croatian and Serbian test set. Manual preprocessing was used. Best scores are boldfaced and split by language.

	set.test		wiki.test		
MTE v4	hr	sr	hr	sr	overall
Lemma	**96.1**	**94.6**	93.9	95.8	**95.1**
POS	95.2	92.3	91.5	90.8	92.5
MSD	86.2	83.4	80.2	81.8	83.1
MTE v5					
Lemma	95.6	94.2	**94.3**	**96.1**	**95.1**
POS	**96.4**	**93.0**	**92.2**	**91.8**	**93.5**
MSD	**86.7**	**84.4**	**80.5**	**82.4**	**83.7**

Table 5: Lemmatization, POS and MSD tagging accuracy on the test sets and overall. Scores are given separately for the two morphosyntactic tagsets used.

the two languages. This data is presented in Table 4. Once again, the language variety is seen to be of no significance to the parsing models. The scores actually alternate in favoring the two languages. SETIMES.HR substantially outperforms HOBS + Sub on the most frequent and arguably the most informative categories, such as predicate and subject (at least 5 LAS points), object (almost 20 LAS points) and attribute (6 points LAS).

4.2 Preprocessing and features

Here we discuss the impact of automatic preprocessing, i.e., lemmatization and MSD tagging on dependency parsing in our test framework. As announced, this discussion deals exclusively with SETIMES.HR. We lemmatize and tag the test samples by using freely available state-of-the-art models for Croatian and Serbian (Agić et al., 2013), parse them using our best SETIMES.HR model and observe LAS and UAS. Preprocessing performance is given in Table 5

	set.test		wiki.test		
LAS	hr	sr	hr	sr	overall
HOBS	57.2	55.9	49.9	51.0	53.8
HOBS + Sub	65.2	62.5	56.7	58.0	60.9
SETIMES.HR	**73.4**	**70.4**	**65.3**	**67.4**	**69.4**
UAS					
HOBS	71.6	71.8	67.4	69.0	70.1
HOBS + Sub	76.2	74.4	71.8	72.5	73.9
SETIMES.HR	**79.4**	**76.9**	**75.2**	**77.8**	**77.4**

Table 6: Parsing accuracy (LAS, UAS) with automatic preprocessing

as a reference point while, more importantly, the dependency parsing scores are given in Table 6. The second plot of Figure 1 provides the learning curves for the automatically preprocessed test sets.

Table 6 scores are easily elaborated using the previously discussed scores with manual, i.e., gold or perfect preprocessing. Namely, the impact of differences between manual and automatic preprocessing on parsing quality basically amounts to a very simple formula: LAS is reduced by 3-4 points and UAS by 2 points when introducing preprocessing noise by automatic lemmatization and tagging. This observation is valid across the languages and domains of our test set and thus applies generally. Keeping in mind the more complex prospective NLP systems for Croatian and Serbian, we consider this fact to be very favorable as the observed 16% error rate in full MSD tagging, 5-6% for POS and lemmatization, amounts for a significantly smaller decrease in parsing quality as quantified by LAS and UAS.

To further support this observation, we conducted an experiment with purposely corrupting lemmatization and tagging. In this, as previously for learning curves, we use the single merged test sample. For lemmatization, we randomly drop lemmas from the manually annotated test sample, replacing them with empty features.[11] For MSD tagging, we implement two procedures. The first is identical with the one for lemmatization, while in the second we replace the valid tag with a randomly selected Croatian tag from the full MTE v5 morphosyntactic tagset. For each

[11] In terms of the CoNLL-X format, we simply replace the valid entry from the LEMMA field by an underscore.

	Croatian		Serbian	
Features	LAS	UAS	LAS	UAS
Adjective				
Type	74.3	80.7	**74.6**	**81.2**
Degree	74.3	80.7	73.7	80.2
Gender	74.1	80.7	**74.5**	**81.0**
Number	74.5	**81.0**	**74.3**	**80.8**
Case	**75.0**	**81.5**	**74.4**	**81.1**
Noun				
Type	74.3	80.8	72.9	80.0
Gender	74.4	80.8	74.1	**80.7**
Number	74.1	80.7	74.0	**80.7**
Case	73.3	81.0	72.3	80.0
Verb				
Type	**74.6**	**81.3**	**74.3**	**80.8**
Form	74.3	80.9	**74.3**	**81.0**
Person	74.3	**81.0**	73.5	80.0
Number	74.4	80.8	74.1	80.6
Gender	74.4	80.8	**74.4**	**81.0**
Full feature set	74.5	80.9	74.1	80.6

Table 7: Impact of morphosyntactic feature exclusion on parsing. Improvements boldfaced and split by language.

of these scenarios, we provide 11 test sets: stepping by 10% of removals or random insertions, from 0% to 100% preprocessing accuracy. The results are plotted in Figure 2. Evidently, lemmatization is of no influence to dependency parsing using our model. This is an important observation to consider in, e.g., the future tasks of parsing large web corpora of Croatian and Serbian. The large impact of morphosyntactic tagging, i.e., morphosyntactic features on parsing is also evident from the figure. It is also supported by previous research in parsing using SETIMES.HR (Agić and Merkler, 2013), where a significant bias towards MSD-based parsing models was found over the POS-only-based models. Tag removal and tag randomization appear to induce a very similar effect of near-linear functional dependency between tagging and parsing. We note that this is not entirely supported by our realistic preprocessing test scenario. It is purely due to the fact that our noise introduction procedure does not relate to the modus in which the stochastic tagger errs in processing unseen text. Namely, MSD tagging errors

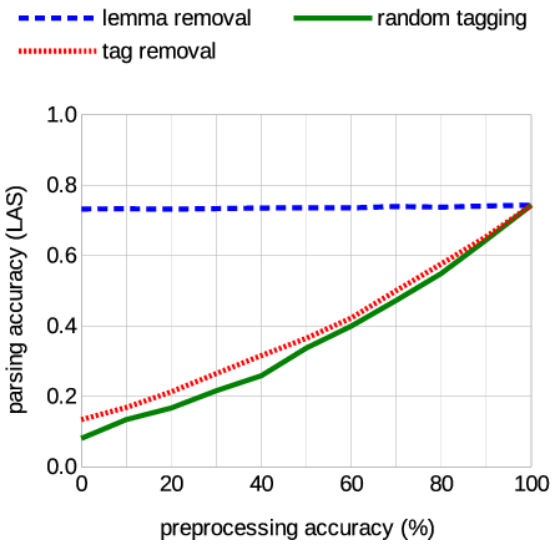

Figure 2: Overall SETIMES.HR parsing accuracy in relation with lemmatization and morphosyntactic tagging

tend to occur on certain morphosyntactic features, corrupting these much more often than entire tags. Thus, even when it yields a feature error, the tagger still provides the parser with other valid features to work with. This consideration of MSD features, in pair with the following set of results, sketches our plans for further research.

Following the previous note on MSD tagset and features, we also implemented a simple experiment in feature weight assessment. In it, we used the SETIMES.HR treebank with full MTE v5 tagset and created from it several instances, each with its own reduced MTE v5 tagset. Each reduction was defined by dropping one MSD feature from one part of speech. More precisely, we dropped all MSD features of adjectives (5 features), nouns (4) and verbs (5). This amounted at 14 different MTE v5 reductions. We trained 14 parsing models using SETIMES.HR with the reduced tagsets and tested them on the test samples merged by language and implementing the respective tagset reductions.

The results are given in Table 7. Most notably, we observed an increase in parsing accuracy when dropping adjective case and verb type. The most substantial decrease occurred with the removal of noun case, indicating the importance of this feature in parsing the two languages. We consider the adjective case removal gain an important observation for future work, as adjectives are the most difficultly

	Adv	Ap	Atr	Atv	Aux	Co	Elp	Obj	Oth	Pnom	Pred	Prep	Punc	Sb	Sub
Adv		0	15	1	0	2	2	5	13	2	1	3	0	2	2
Ap	1		10	0	0	0	2	3	0	1	0	0	0	5	0
Atr	23	9		6	1	0	14	23	3	3	3	0	0	25	2
Atv	0	1	6		0	0	0	0	0	1	26	0	0	1	0
Aux	0	0	0	0		1	0	0	0	0	28	0	0	0	1
Co	0	0	1	0	0		0	0	5	0	0	2	11	0	0
Elp	1	2	12	0	0	0		0	4	3	2	0	0	4	0
Obj	6	3	16	3	0	0	1		0	1	1	0	0	2	0
Oth	14	4	3	0	0	12	1	1		0	0	1	0	1	24
Pnom	3	0	8	0	0	0	3	0	0		24	1	0	3	0
Pred	1	0	2	5	26	0	0	1	1	23		0	0	0	0
Prep	1	0	0	0	1	1	0	0	2	0	0		0	0	0
Punc	0	0	0	0	0	17	0	0	0	0	0	0		1	1
Sb	2	11	26	1	0	0	5	1	4	4	1	0	0		1
Sub	1	0	0	0	0	0	0	0	2	0	0	0	0	0	

Table 8: Confusion matrices for LAS (Croatian: bottom left, Serbian: top right)

tagged category for Croatian and Serbian.

4.3 Error analysis

Here we provide a brief insight to the error instances. We discuss LAS errors for both languages, i.e., instances of invalid head attachments paired with tag misassignments. These are given in Table 8 in the form of two confusion matrices for LAS.

We isolate several clusters of errors with shared linguistic properties. Firstly, the subject-attribute-apposition group (*Sb-Atr-Ap*), in which we find the error instances to be closely related to the order of attachment and assignment in multi-word units representing foreign personal names, titles or functions and occupations of persons. Next, the attribute-adverb-object group (*Atr-Adv-Obj*) expectedly appears as these are inherently ambiguous categories.[12] The predicate-nominal-auxiliary group of errors (*Pred-Pnom-Aux*) reflects the interaction of MSD annotation choices and syntactic annotation principles, as participes are MSD-tagged as adjectives, thus confusing the parser in predicate annotation. Moreover, SETIMES.HR has documented issues with consistency in complex predicate annotation that seek resolution and negatively influence the parsing scores. Lastly, the only error group substantially reflecting the language difference is the one involving predicates and predicate complements (*Pred-Atv*), as it appears only in the Serbian confusions. Namely, the infinitive predicate complement is frequent in Croatian and non-existent in Serbian. Infinitives in Serbian only appear for the future tense paired with auxiliary verbs, confusing the parser to annotate these infinitives as predicate complements as observed in the Croatian training data.

5 Conclusions and future work

We have described an experiment with dependency parsing of two closely related and under-resourced languages, Croatian and Serbian, by using parsing models trained on Croatian treebanks. We investigated three different parsing formalisms, the effects of lemmatization, morphosyntactic tagging and feature selection on parsing quality for both languages. We observed state-of-the-art parsing scores. All resources used in the experiment are made publicly available under a permissive license.[13]

The results of this experiment sketch the path for our future research. Experiments with syntactic projection between Croatian and Serbian are not feasible given the negligible differences in the observed scores. In contrast, domain adaptation for parsing the two languages should be investigated given the observed accuracy decrease when moving from newspaper text to Wikipedia. We have already initiated further enlargements of the SETIMES.HR treebank and the test sets with Croatian data from other domains. Experiments with newer and more advanced dependency parsers (Koo and Collins, 2010; Bohnet and Nivre, 2012; Zhang and McDonald, 2012; Martins et al., 2013) should be conducted to provide up-to-date scores.

We are currently experimenting with morphosyntactic tagset design for improved dependency parsing of Croatian and Serbian. We aim at finding the optimal tagset by closely investigating morphosyntactic feature influences and dependencies.

[12] PDT, e.g., has an *AtrAdv*, *AdvAtr*, *AtrObj* and *ObjAtr* ambiguity classes to address this. However, the sum of their frequencies in HOBS is negligibly small ($< 0.03\%$).

[13] http://nlp.ffzg.hr/

References

Ž. Agić. 2012. K-Best Spanning Tree Dependency Parsing With Verb Valency Lexicon Reranking. In: *Proceedings of COLING 2012: Posters*, pp. 1–12. COLING 2012 Organizing Committee.

Ž. Agić, D. Merkler. 2013. Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. In: *Text, Speech and Dialogue. Lecture Notes in Computer Science*, 8082:560–567. Springer.

Ž. Agić, N. Ljubešić, D. Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In: *Proceedings of BSNLP 2013*. ACL.

B. Bekavac, S. Seljan, I. Simeon. 2008. Corpus-Based Comparison of Contemporary Croatian, Serbian and Bosnian. In: *Proceedings of FASSBL 2008*, pp. 33–39. Croatian Language Technologies Society.

D. Berović, Ž. Agić, M. Tadić. 2012. Croatian Dependency Treebank: Recent Development and Initial Experiments. In: *Proceedings of LREC 2012*, pp. 1902–1906. ELRA.

A. Böhmová, J. Hajič, E. Hajičova, B. Hladká. 2003. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: *Treebanks: Building and Using Parsed Corpora*. Springer.

B. Bohnet, J. Nivre. 2012. A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In: *Proceedings of EMNLP-CoNLL 2012*, pp. 1455–1465. ACL.

S. Buchholz, E. Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In: *Proceedings of CoNLL-X*, pp. 149–164. ACL.

T. Erjavec. 2012. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46 (1), 131–142. Springer.

Y. Goldberg, M. Elhadad. 2009. Hebrew Dependency Parsing: Initial Results. In: *Proceedings of IWPT 2009*, pp. 129–133. ACL.

K. Hall. 2007. K-Best Spanning Tree Parsing. In: *Proceedings of ACL 2007*, pp. 392–399. ACL.

T. Koo, M. Collins. 2010. Efficient Third-Order Dependency Parsers. In: *Proceedings of ACL 2010*, pp. 1–11. ACL.

M. Martinović. 2008. Transfer Of Natural Language Processing Technology: Experiments, Possibilities and Limitations – Case Study: English to Serbian. *Infotheca – Journal of Informatics and Librarianship*, 9 (1-2):11–20.

A. Martins, M. Almeida, N. Smith. 2013. Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers. In: *Proceedings of ACL 2013*. ACL.

R. McDonald, K. Lerman, F. Pereira. 2006. Multilingual Dependency Parsing With a Two-Stage Discriminative Parser. In: *Proceedings of CoNLL-X*, pp. 216–220. ACL.

D. Merkler, Ž. Agić, A. Agić. 2013. Babel Treebank of Public Messages in Croatian. In: *Proceedings of CILC 2013. Proceedia – Social and Behavioral Sciences*, in press. Elsevier.

E. Mohamed. 2011. The Effect of Automatic Tokenization, Vocalization, Stemming, and POS Tagging on Arabic Dependency Parsing. In: *Proceedings of CoNLL 2011*, pp. 10–18. ACL.

G. Nenadić. 2000. Local Grammars and Parsing Coordination of Nouns in Serbo-Croatian. In: *Text, Speech and Dialogue. Lecture Notes in Computer Science*, 1902:57–62. Springer.

G. Nenadić, I. Spasić, S. Ananiadou. 2003. Morphosyntactic Clues for Terminological Processing in Serbian. In: *Proceedings of the EACL Workshop on Morphological Processing of Slavic Languages*, pp. 79–86. ACL.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, D. Yuret. The CoNLL 2007 Shared Task on Dependency Parsing. In: *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pp. 915–932. ACL.

M. Silberztein. 2004. NooJ : An Object-Oriented Approach. In: *INTEX pour la Linguistique et le Traitement Automatique des Langue*, pp. 359–369. Presses Universitaires de Franche-Comté.

J. Silić, I. Pranjković. 2005. Gramatika hrvatskoga jezika za gimnazije i visoka učilišta. Školska knjiga, Zagreb.

A. Søgaard. 2013. Semi-Supervised Learning and Domain Adaptation for NLP. Morgan & Claypool Publishers.

Ž. Stanojčić, Lj. Popović. 2008. Gramatika srpskog jezika: za gimnazije i srednje škole. Zavod za udžbenike i nastavna sredstva, Beograd.

M. Tadić. 2007. Building the Croatian Dependency Treebank: The Initial Stages. *Suvremena lingvistika*, 63 (1), 85–92. Hrvatsko filološko društvo.

M. Tadić, D. Brozović-Rončević, A. Kapetanović. 2012. The Croatian Language in the Digital Age. *META-NET White Paper Series*. Springer.

M. Tadić, T. Váradi. 2012. Central and South-East European Resources in META-SHARE. In: *Proceedings of COLING 2012: Demonstration Papers*, pp. 431–438. COLING 2012 Organizing Committee.

D. Vitas, C. Krstev, I. Obradović, Lj. Popović, G. Pavlović-Lažetić. 2003. An Overview of Resources and Basic Tools for Processing of Serbian Written Texts. In: *Proceedings of the Workshop on Balkan Language Resources, First Balkan Conference in Informatics*.

D. Vitas, Lj. Popović, C. Krstev, I. Obradović, G. Pavlović-Lažetić, M. Stanojević. 2012. The Serbian Language in the Digital Age. *META-NET White Paper Series*. Springer.

K. Vučković, M. Tadić, Z. Dovedan. 2008. Rule-Based Chunker for Croatian. In: *Proceedings of LREC 2008*, pp. 2544–2549. ELRA.

D. Yarowsky, G. Ngai, Richard Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In: *Proceedings of HLT 2001*, pp. 1–8. ACL.

H. Zhang, R. McDonald. 2012. Generalized Higher-Order Dependency Parsing With Cube Pruning. In: *Proceedings of EMNLP 2012*. ACL.