Croatian Dependency Treebank 2.0: New Annotation Guidelines for Improved Parsing

Željko Agić,[⋆] Daša Berović,[†] Danijela Merkler,[†] Marko Tadić[‡]
[⋆]Linguistics Department, University of Potsdam, Germany
[‡]Department of Linguistics, Faculty of Humanities and Social Sciences, [†]University of Zagreb, Croatia
zagic@uni-potsdam.de, {dberovic, dmerkler, mtadic}@ffzg.hr

Abstract

We present a new version of the Croatian Dependency Treebank. It constitutes a slight departure from the previously closely observed Prague Dependency Treebank syntactic layer annotation guidelines as we introduce a new subset of syntactic tags on top of the existing tagset. These new tags are used in explicit annotation of subordinate clauses via subordinate conjunctions. Introducing the new annotation to Croatian Dependency Treebank, we also modify head attachment rules addressing subordinate conjunctions and subordinate clause predicates. In an experiment with data-driven dependency parsing, we show that implementing these new annotation guidelines leeds to a statistically significant improvement in parsing accuracy. We also observe a substantial improvement in inter-annotator agreement, facilitating more consistent annotation in further treebank development.

Keywords: dependency treebank, dependency parsing, Croatian language

1. Introduction

Croatian Dependency Treebank (Tadić, 2007) (HOBS further in the text) is built according to the model developed for the Prague Dependency Treebank (Böhmová et al., 2003) (PDT). The fact that these two morphologically rich Slavic languages have similar syntactic structures enabled the adaptation of Czech syntactic formalism for Croatian. However, not all language phenomena in Croatian are identical to those in Czech. In the course of annotating the first version of HOBS on the syntactic (or, observing the PDT terminology, analytical) level, we have encountered a number of issues with the annotation of subordinate clauses. The aim of this work is to recognize the differences in treatment of subordinate clauses in Croatian and Czech and to suggest a new approach to annotating HOBS according to the observed differences.

Accounting for these annotation inconsistencies in HOBS, we propose a new approach to subordinate clause annotation — developed specifically for Croatian and thus different from the one used in PDT — considering that Croatian grammars and dictionaries treat syntactic conjunctions different then Czech grammars and dictionaries. We put special emphasis on the attribute clause and the adverbial clause. Firstly, this is due to the fact that the biggest inconsistencies between HOBS and PDT occur specifically in the annotation of attribute clauses. Secondly, adverbial clauses in Croatian have a rich classification, implying the possible importance of including this classification to HOBS as well. We propose a subset of syntactic tags for all different types of adverbial clauses. We derive the proposed tag subset from the analysis of clauses in HOBS and by consulting Croatian grammars.

We manually convert HOBS with respect to the suggested set of new syntactic and respective head attachment rules for subordinate clause annotation. We then use both versions of HOBS — the one with implicit and the one with explicit subordinate clause annotation – in an experiment with data-driven graph-based dependency parsing. The proposed annotation scheme, the resulting treebank and the experiment results are discussed in the following sections. First, we describe the annotation approach in the PDT-conformant HOBS, which is followed by a detailed account on its adaptation towards explicit annotation of subordinate clause predicates. Second, the adaptation is implemented, in turn creating a new version of HOBS for which we observe an increase in inter-annotator agreement over the PDT-conformant version. Finally, we use both versions in an experiment involving a data-driven dependency parser to show substantial improvements in parsing accuracy. We conclude by sketching possible directions for future research in Croatian dependency treebanking and data-driven parsing.

2. Subordinate clause annotation in HOBS

In this section, we elaborate on the drawbacks of PDT-style annotation of subordinate clauses in HOBS and propose an approach to explicit annotation of subordinate clause predicates via syntactic conjunctions.

2.1. PDT-style guidelines

On the syntactic level of corpus annotation (for PDT and HOBS) dependency relations between sentence elements are shown. Besides, every sentence element is assigned with a label denoting its syntactic function. Syntactic structure of the sentence is represented by an acyclic graph, i.e., by a parse tree. Every node of the tree is labeled with one of the 28 basic syntactic tags which should reflect the syntactic role of each node in the sentence (Hajič et al., 2001). However, some of the syntactic tags are based on semantic criteria, not exclusively on syntactic criteria. These functions are assigned to sentence elements which cannot be annotated with syntactic tags for traditional syntactic elements, e.g., subject or predicate, using syntactic tags AuxO (redundant or emotional item) and AuxZ (emphasizing word). The formalism of dependency grammar was a guide for the representation of the links between heads and subordinate elements in sentences. In the sentence structure, the predicate (i.e., verb) in the subordinate clause is subordinated to the verb in the main clause. Every subordinate clause takes its place as one of the syntactic elements in the main clause. Predicate in the subordinate clause is annotated with the syntactic tag of a syntactic element whose place it takes in the main clause (e.g., if the subordinate clause was subject clause, its predicate is assigned with the syntactic tag Sb). Subordinate clauses can be introduced to the main clause in two ways. One is direct introduction without any of the syntactic conjunctions, in which the predicate in the subordinate clause depends directly on the predicate in the main clause. In the examples with syntactic conjunctions there are differences in the annotation in HOBS and PDT, due to the different interpretations of conjunctions as parts of speech and conjunctions as syntactic functions in the Croatian and Czech grammars.

In PDT all subordinate clauses with conjunctions are introduced to the main clause in the same way. The principle is based on the classification of syntactic conjunctions. Besides real conjunctions (as a part of speech), a syntactic conjunction in Czech can also be a pronoun or an adverb. On the syntactic level, only real conjunctions can be annotated with syntactic tag AuxC. In the tree, they can mediate between the predicate in the main clause and the predicate in the subordinate clause. Conjunction as a part of speech cannot be one of the elements of syntactic structure and only conjunctions as a part of speech are considered as real syntactic conjunctions. Other words that introduce subordinate clauses are not considered to be conjunctions and they are introduced to the subordinate clause as elements of its syntactic structure. In such examples, the predicate in the subordinate clause is directly dependent on the predicate in the main clause, and conjunction word is annotated with the syntactic tag of a syntactic element whose place it takes in the sentence.

The principle of annotating subordinate clauses in HOBS is different than the principle of annotating in PDT. All subject, object, predicate and adverbial clauses are introduced to the main clause in the same way. This is the pattern for the annotation of all subordinate clauses – all conjunctional words (conjunctions, pronouns and adverbs) are annotated as syntactic conjunctions and they are intermediaries between the predicate in the subordinate and the predicate in the main clause – except attribute clauses. There are also some very rare examples in which the predicate in subordinate clause depends directly on the predicate in the main clause because the conjunctional word is not present in the sentence.

Annotation of attribute clauses in HOBS follows the pattern of annotation in PDT. In examples with conjunctions as a part of speech, predicate in the subordinate clause depends on the conjunction (see Figure 1a). But in the examples in which syntactic conjunction was an adverb, pronoun or prepositional phrase, conjunction becomes one of the elements of syntactic structure in the subordinate clause and the predicate in the subordinate clause depends directly on the predicate in the main clause.

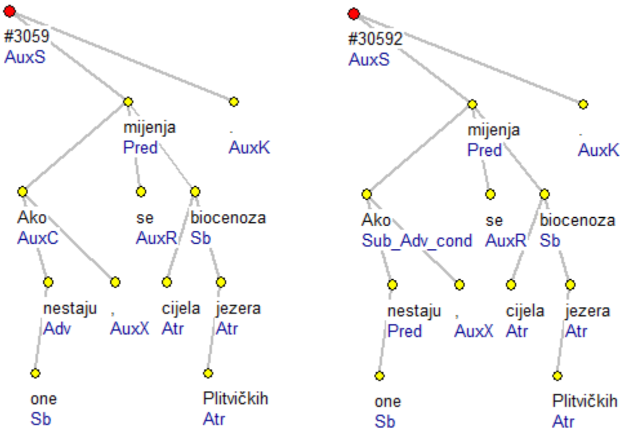

Figure 1: An example of a subordinate clause in the old version of HOBS and its adaptation for the new version. Note the introduction of tag Sub_Adv_cond instead of AuxC for the conjunction *Ako* and the explicite annotation of the subordinate clause predicate *nestaju* instead of the implicit Adv tag. (hr: *Ako one nestaju, mijenja se cijela biocenoza Plitvickih jezera.*, en: *If they are receding, it changes the entire biocenosis of the Plitvice lakes.*)

2.2. Annotation proposal

In Croatian, all types of subordinate clauses are treated in the same way. Subordinate clauses take place as one of the syntactic elements in the main clause, which have a clear syntactic function. Besides, all subordinate clauses are introduced to the main clause in several different ways: with conjunction (as a part of speech), pronoun, adverb or prepositional phrase. In the representation of the sentence, conjunction has to be annotated as conjunction – it cannot be a syntactic element in the subordinate clause because it introduces a subordinate clause to the main clause. In (Raguž, 1997) it is confirmed that words that link clauses have the syntactic function of conjunction: ”That is why conjunctions are a special part of speech, although in the strict sense they are not parts of speech because that is their syntactical function, similar to adjectives being assigned attributes on the syntactic level. Some adverbs or pronouns become conjunctions.” We propose that all conjunctions in the tree should depend on the predicate in the main clause, and that the predicate in the subordinate clause should depend on the conjunction. There is one more reason for a new approach to the annotation of subordinate clauses. In the current annotation, the predicate in the subordinate clause is assigned with the syntactic tag of the sentence element whose place in the sentence structure it takes. From an information extraction point of view and regarding the consistency of the annotation, we consider this approach to be somewhat insufficient with regards to the encoded information on the syntactic structure of the sentence. Predicate in the subordinate clause — in the same way as the predicate in the main clause — should be annotated with the syntactic tag Pred, because that is its syntactic function. In order not to lose information of the type of clause which is introduced to the main clause, we propose that syntactic tag of the conjunction is assigned with a label from the subset of labels which would give the information of the subordinate clause

Clause type	Syntactic tag
attribute	Sub_Atr
adverbial	Sub_Adv
object	Sub_Obj
predicate	Sub_Pred
subject	Sub_Sb

Table 1: The new syntactic sub-tagset for subordinate clause annotation in HOBS

Conjunction	Adverbial clauses
kako	modal, causative, final, consequential, temporal
da	consequential, final, conditional, concessive, (causative)
što	causative, modal, temporal
kad, kada	temporal, causative, conditional

Table 2: The most frequent syntactic conjunctions related to the corresponding adverbial clauses

type. The conjunction introducing the subordinate clause would have the syntactic tag Sub (and not AuxC) in order to indicate that it is the subordinated clause and in order to set a correlation with coordinated clauses which are introduced with label Coord. While introducing subordinate clauses, the suggested label Sub would get a sub-label for the type of the subordinate clause (see Table 1).
Development of the subsets of labels and conjunction annotation by an universal syntactic tag reduces the differences between various representations of subordinate clauses, and all sentence elements of the subordinate clauses get the label for the syntactic function they perform (see Figure 1b). Besides the previously stated, further in the text we propose a new annotation scheme for adverbial clauses, considering their significant sub-classification in Croatian. Development of the subsets of labels for the detection of subordinate clause types and types of adverbial clauses enables stratified sentence representation. Information obtained from corpus annotated in that way can be reduced to the minimum or increased to the maximum, depending on the required abstraction level.

2.2.1. Attribute clauses

Attribute clauses are the most frequent type of relative clauses, which are in turn the most frequent type of the subordinate clauses in Croatian. They are most often described as subordinate clauses which specify a certain nominal word in the main clause. Relative pronoun *koji* and relative adverb (or conjunction) *što* most frequently have the syntactic function of conjunction in attribute clauses. It is said that the main difference in determination and in the way of introduction of attribute clause in PDT and HOBS is based on word type, i.e., on the syntactic function of the word that introduces a subordinate clause. In Croatian grammars and dictionaries,[1] these two syntactic conjunctions are not unambiguously determined. In both grammars, *koji* is independently determined as a relative pronoun, but it is clearly noted that it has the syntactic function of conjunction in relative attribute clauses. On the contrary, in Croatian dictionaries — which should describe all functions for a certain word — there are different explanations: one dictionary (Anić, 2003) describes *koji* only as a (relative) pronoun, and another (Šonje, 2000) separates *koji* as a pronoun and *koji* as a conjunction. Even in consulted grammars, *što* is not unambiguously determined: in (Barić et al., 1995) it is described as a relative adverb, and in (Silić et al., 2005) as a relative conjunction. But in both grammars it is said that ”relative adverb *što* substitutes relative pronoun *koji* in nominative case in all three genders and both numbers” (Barić et al., 1995), that is — more specifically — ”relative conjunction *što*, which occurs with unstressed forms of personal pronouns in attribute clause, is always replaceable with the pronoun *koji*” (Silić et al., 2005). It is confirmed in (Pranjković, 1986) that relative conjunction *što*[2] is less frequent than relative *koji*, because it is somewhat more complex, considering it is determined by certain grammatical and semantic characteristics, but *što* is used also as a sort of stylistic backup for relative *koji*, which is recommended considering the frequency and recursion of relative clauses that may lead to accumulation of relative *koji*. Furthermore, it is said in (Šonje, 2000) that the clauses with relative *što* are ”equivalent with clauses with relative pronoun *koji*”. Considering the described possibility of mutual replacement of relative *koji* and relative *što* — like in example from (Silić et al., 2005): *Pjesma **koje** si se sjetio još se pjeva.* (Song which you remembered is still sung.) and *Pjesma **što** si **je** se sjetio još se pjeva.* (Song that you remembered is still sung.) — and having in mind that, in examples like this, relative *što* is defined as a relative conjunction which introduces relative attribute clause, we think that relative *koji* in the same examples (see Figure 2b) should be equally defined and equally annotated as relative *što* in description of the dependent structure of clauses (former processing of attribute clauses in HOBS is presented in Figure 2a).

Adverbial clause	Sub_Adv
local	Sub_Adv_loc
temporal	Sub_Adv_temp
modal	Sub_Adv_mod
causative	Sub_Adv_caus
consequential	Sub_Adv_cons
final	Sub_Adv_fin
conditional	Sub_Adv_cond
concessive	Sub_Adv_cons

Table 3: Additional syntactic sub-tags for adverbial clause subclassification

[1] For this paper, two Croatian grammars ((Barić et al., 1995) and (Silić et al., 2005)) and two Croatian dictionaries ((Anić, 2003) and (Šonje, 2000)) were consulted.

[2] It is strictly apparted from relative pronoun *što* just because the conjunctional type is always replaceable with *koji*, but the pronominal type never is.

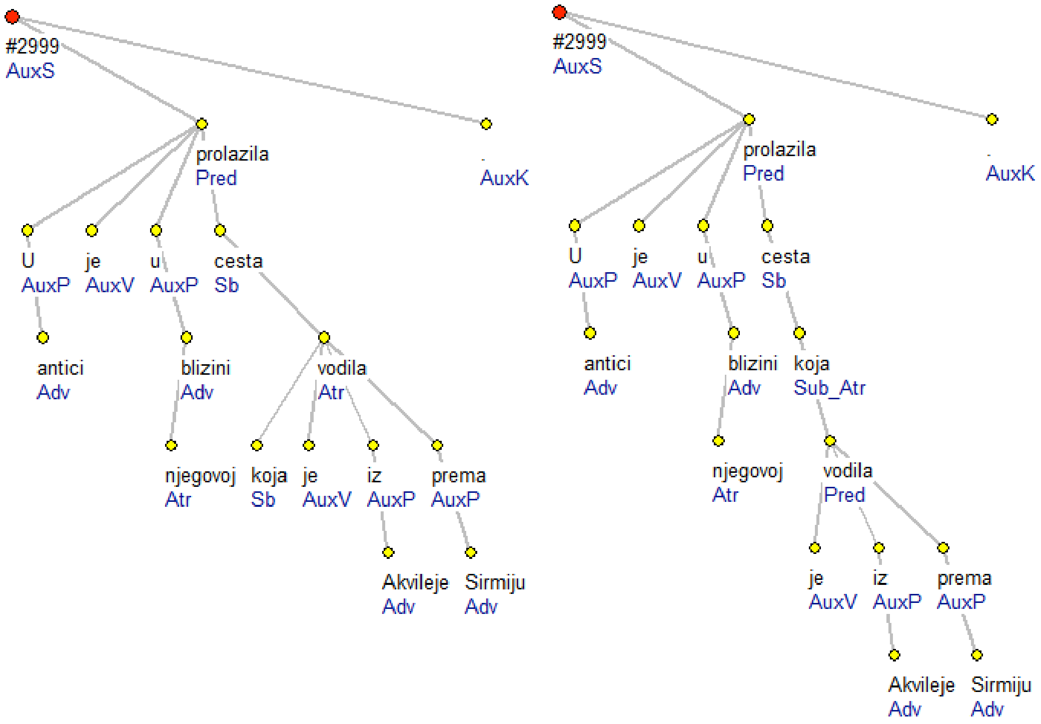

Figure 2: (a) Attribute clause in old HOBS and (b) Proposal for attribute clause annotation (hr: *U antici je u njegovoj blizini prolazila cesta koja je vodila iz Akvileje prema Sirmiju.*, en: *In the classical period, a road that lead from Aquileia to Sirmium passed near it.*)

2.2.2. Adverbial clauses

Adverbial clauses are a separate type of subordinate clauses in HOBS, with no further classification of different types of adverbial clauses. In consulted Croatian grammars, adverbial clauses are distributed — depending on which type of adverb in the main clause they stand for — in eight or nine[3] different types of clauses. In this paper, we concentrate on the types of adverbial clauses which are described in both grammars. These are: local, temporal, modal, causative, consequential, final, conditional and concessive clauses. All types of adverbial clauses can be introduced to the main clause with different syntactic conjunctions. Some of them introduce different types of clauses to the main clause — three, four or even five different types. Table 2 shows four of the most frequent syntactic conjunctions of adverbial clauses and the type of subordinate clause they can introduce to the main clause.

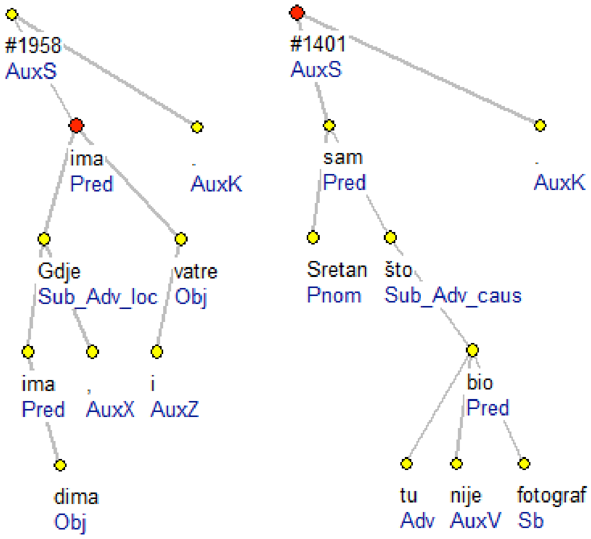

Figure 3: Proposal for (a) causative clause annotation (hr: *Gdje ima dima, ima i vatre.*, en: *Where there's smoke, there's fire.*) and (b) local clause annotation (hr: *Sretan sam što tu nije bio fotograf.*, en: *I'm happy a fotographer wasn't there.*)

It shows that the conjunction *što* is among the most frequent syntactic conjunctions of adverbial clauses, and it is also — as it was previously described — one of the two most frequent conjunctions for introduction of relative attribute clauses. Therefore, *što* is not only the syntactic conjunction for different types of adverbial clauses, but it is a conjunction for two basic types of subordinate clauses — adverbial and attribute — so we can say its conjunctional multifuncionality is even more substantial. We consider that — besides these two basic differences in the usage of conjunction *što* — the difference of its usage in adverbial clauses, as well as the usage of other multifunctional syntactic conjunctions, should be distinguished and annotated (see Figure 3). These additional labels would not be new independent labels, but rather a sort of expansion or amendment to the previously proposed label Sub_Adv. The proposal of such annotation is given in Table 3. Although, on the one hand, it could seem that this kind of label expansion would be a complication of the annotation process and an unnec-

[3] One more adverbial clause type, namely, the comparative clause, is identified in (Silić et al., 2005).

Treebank	Tagset	Tokens	MSDs	Tags
HOBS 1.0	basic	117 369	914	27
	full	117 369	914	70
HOBS 2.0	basic	117 369	911	28
	full	117 369	911	81

Table 4: Basic treebank statistics: number of tokens, distinct lemmas and syntactic tags for both versions of HOBS

Treebank	Tagset	LAS	UAS	LA	ϰ(LA)
HOBS 1.0	full	75.01	86.44	81.99	0.810
HOBS 2.0	basic	82.05	89.16	88.83	0.884
	full	78.89	89.16	84.07	0.839

Table 5: Inter-annotator agreement

essary addition of new labels in an already big existing set of syntactic tags, this type of sub-labels would, on the other hand, reduce the complexity of search and, in addition, it would provide more concrete information. In this way, it would be possible, for example, to single out only modal clauses by searching for this specific label (Sub_Adv_mod), or to retain on more abstract levels and search only by basic labels (Sub_Adv or even Sub).

3. Experiments

In this section, we discuss the impact that implementing the new annotation proposal in HOBS has on inter-annotator agreement for manual annotation and on the quality of data-driven dependency parsing of Croatian.

3.1. Annotation consistency

We used the new annotation proposal over the existing PDT-style annotation of HOBS, altering the head attachment and the syntactic tags for every subordinate clause in the treebank. This in turn derived a new version of the treebank: henceforth, we will refer to the PDT-style version as HOBS 1.0 and we dub the new version with the syntactic tagset expansion as HOBS 2.0. Note that we also differentiate between two sub-versions of the each treebank throughout the experiment: the version with the basic syntactic tagset and the version with the full tagset. This follows the PDT-style definition of syntactic tags and subtags in which a syntactic tag is divided into the basic and extended part by an underscore. For example, the syntactic tag Pred_Co denotes a predicate (basic tag: Pred) which participates in a coordination structure with another predicate (extension: Co, full tag: Pred_Co). Further in the text, a reference to HOBS 1.0 or 2.0 with just the basic or the full syntactic tagset will denote this distinction.

The basic statistics for the two versions are given in Table 4. As to the basic counts, the two treebanks expectedly have the same number of sentences (4,626) and tokens (and also types and lemmas). The morphological tag counts differ slightly due to minor error corrections, while the major differences are exhibited by counts and distributions of syntactic tags from the respective syntactic tagsets.

For quantifying the consistency of the new annotation style in comparison with the PDT-style annotation, we calculate the agreement between two expert annotators on a development set of 100 sentences extracted from HOBS 1.0 (full tagset) and HOBS 2.0 (basic and full tagset). The results are displayed in Table 5. We calculated the standard dependency parsing accuracy metrics: labeled and unlabeled attachment score (LAS, UAS) and sequential label attachment (LA). We also used the sequential attachment of labels to calculate Cohen's kappa ϰ(LA) as an indicator of the actual agreement of annotators accounting for agreement by chance. The data in the table clearly indicates that the new annotation guidelines implemented in HOBS 2.0 facilitate easier and more reliable annotation as the improvements are substantial according to all metrics. It is worth noting that even if the full HOBS 2.0 syntactic tagset is 11 tags larger than the HOBS 1.0 tagset, the improvements in both label assignment (LA, LAS) and head attachment (UAS) are consistently large. Drawing from these scores, it is safe to claim that the new annotation guidelines are better suited for syntactic annotation of Croatian text than the PDT guidelines, which are in turn motivated by Czech syntactic analysis. As expected, shrinking the full HOBS 2.0 tagset into its basic version further raises the scores due to tagset simplification. Even if less important, this can still be a useful observation for, e.g., dependency parsing applications which don't require a large and expressive syntactic tagset to operate.

3.2. Dependency parsing

The Croatian Dependency treebank project was first initiated in 2007 (Tadić, 2007). Thus, HOBS was at that time not large enough to be included in the standard benchmarks in data-driven dependency parsing in the field, such as the CoNLL 2006 and 2007 shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007a). In an effort to evaluate and improve standard dependency parsing paradigms on Croatian text from HOBS, a number of research directions were explored prior to this work. Berović et al. (2012) apply a standard transition-based parser MaltParser (Nivre et al., 2007b) to a prototype HOBS with approximately 2,700 sentences to reach 71 LAS points in a tenfold cross-validation scenario. Agić (2012) uses a 3,450 sentence strong HOBS prototype to compare transition-based and graph-based parsing paradigm of the MSTParser generator (McDonald et al., 2005), establishing a strong preference for the latter one, as Croatian exhibits a large quantity of non-projectivity in HOBS (more than 20% at sentence level). Furthermore, (Agić, 2012) suggests a novel method for hybrid graph-based parsing based on parse tree evaluation and reordering using a valency lexicon of Croatian verbs – CROVALLEX (Mikelić Preradović et al., 2009). The improvements amount to an overall LAS score of approximately 77. More recently, Agić and Merkler (2013) compare HOBS with a newly-developed SETimes.HR dependency treebank of Croatian, the latter implementing a simplistic syntactic formalism with only 15 tags and thus aiming at higher dependency parsing performance for tasks in which the syntactic tagset expressivity is not of crucial importance. Agić et al. (2013) build on this comparison by

Treebank	Tagset	LAS	UAS	LA
HOBS 1.0	basic	71.93	79.98	84.65
	full	71.71	80.34	81.75
HOBS 2.0	basic	74.50	81.41	86.87
	full	73.04	81.10	82.85

Table 6: Overall parsing accuracy for the two versions of HOBS with the basic and the full syntactic tagset

involving HOBS and SETimes.HR in an experiment with direct lexicalized transfer parsing. In addition, Merkler et al. (2013) attempt to directly apply these models to non-standard Croatian text. However, as we don't address the relationship between syntactic formalism expressivity and dependency parsing accuracy in this research, we only provide comparison between HOBS versions 1.0 and 2.0 as to the underlying differences in annotation guidelines and their downstream effects. Both versions of HOBS are compliant with the MTE v4 morphosyntactic tagset specification (Erjavec, 2012), enabling direct comparability.

For the experiment, we use the graph-based MSTParser generator system[4] (McDonald et al., 2005). With a number of "second- and third-generation" dependency parsing systems now publicly available (Bohnet, 2010; Bohnet et al., 2013) that consistently outperform the standard parsers such as MaltParser and MSTParser, our choice is motivated by backward compatibility with previous research in Croatian dependency parsing and by exploiting the available MSTParser models for HOBS 1.0. Also, this being a syntactic annotation paradigm comparison, we don't explicitly aim at reaching state-of-the-art scores in terms of measures such as LAS and UAS, but rather at designing an optimal syntactic tagset in terms of joint optimization of annotation quality, tagset expressivity and parsing accuracy.

We create a standard tenfold cross-validation parsing experiment with a 9:1 treebank division between the training and the testing set. We use approximate randomization for statistical significance testing where applicable. The first set of results, i.e., the overall parsing accuracy is given in Table 6. The table reveals a strong preference for the HOBS 2.0 treebank across the scores, as even the models with the full HOBS 2.0 syntactic tagset (81 tag) significantly outperform both the basic HOBS 1.0 (27 tags) and full HOBS 1.0 (70 tags) tagset models. The difference is maintained for all three parsing accuracy metrics. This indicates that the search for an optimal syntactic tagset does not necessarily have to be reduced to a simple inverse proportionality between the tagset size and the performance of the parser, even if this rule expectedly holds in most cases (Mille et al., 2012; Agić and Merkler, 2013).

In Table 7, the parsing scores are grouped by 10 main (and most frequent) syntactic tags. Together with the scores, the table shows test set frequencies as an indicator of a specific tag's impact on the overall scores. HOBS 2.0 models outperform HOBS 1.0 by a very large margin for the most important syntactic tags: predicate (80.69 vs. 65.89 in LAS), subject (73.99 vs. 68.85) and object (70.06 vs. 62.81). The differences in favor of the new annotation scheme of HOBS 2.0 hold for most other syntactic tags, with exceptions being underrepresented in the test set. We consider the accuracy gain on the basic syntactic categories to be a very important observation. We particularly note that the number of word forms annotated as predicates has doubled moving from HOBS 1.0 (4.76%) to HOBS 2.0 (9.99%) due to explicit annotation of subordinate clause predicates, while the increase in head attachment and label assignment for predicates has gone up by 14.8 LAS points. We believe that implications of better identification of clause predicates for information extraction tasks that build on dependency parsing are very favorable, but this should be further verified by downstream evaluation. It should also be noted that the labeled attachment score for subordinate conjunction (Sub) measures favorably against the overall accuracy for HOBS 2.0, but the head attachment (UAS) still needs improvement, which could possibly be addressed by a closer inspection of error properties.

	HOBS 1.0 basic			HOBS 2.0 basic		
Tag	LAS	UAS	pct	LAS	UAS	pct
Adv	65.88	84.81	9.98	**68.33**	**88.33**	8.99
Apos	**38.10**	**47.62**	0.64	36.84	42.11	0.64
Atr	81.61	88.29	28.7	**83.06**	**89.18**	25.8
Coord	48.21	49.23	4.15	**56.85**	**59.39**	4.18
Obj	62.81	79.40	8.39	**70.06**	**87.65**	6.53
Pnom	58.73	**80.95**	1.51	**60.61**	77.27	1.74
Pred	65.89	72.87	4.76	**80.69**	**82.19**	9.29
AuxP	69.85	70.50	9.28	**71.54**	**71.94**	9.99
Sb	68.85	81.26	7.84	**73.99**	**82.37**	7.01
Sub	–	–	–	72.91	73.89	4.04

Table 7: Parsing accuracy and test set frequencies for matching syntactic functions in the two versions of HOBS with basic syntactic tags only

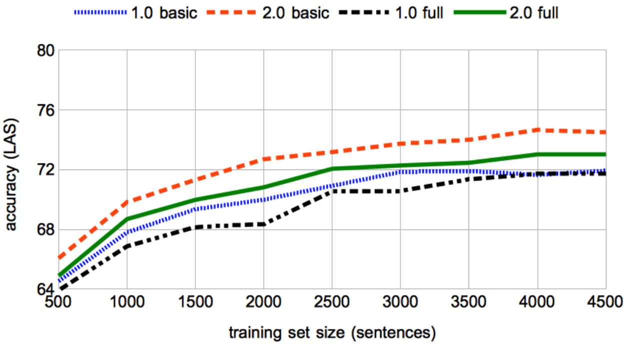

Figure 4: Learning curves (LAS) for the two versions of HOBS with the basic and the full syntactic tagset

The learning rate experiment involved splitting the treebanks into 9 incrementally enlarged subsets ranging from 500 to 4,500 sentences, training and testing the models. Figure 4 displays the LAS learning curves for HOBS 1.0 and 2.0 with the basic and the full syntactic tagset. Learning rates are comparable, with a clear distinction between the higher-scoring HOBS 2.0 and the lower-scoring HOBS

[4] http://sourceforge.net/projects/mstparser/

1.0 models' learning curves. Due to smaller size, the basic tagset learning curves also top the respective full tagset learning curves.

4. Conclusions and future work

In this contribution, we presented the new version of the Croatian Dependency Treebank – HOBS 2.0. It implements an extension of the Prague Dependency Treebank syntactic layer annotation formalism, that was closely observed in the previous version of HOBS. The extension deals with explicit annotation of predicates in subordinate clauses and it introduces a set of new syntactic tags for the annotation of syntactic subordinating conjunctions. We compared the newly-developed HOBS 2.0 with the previous edition of the treebank (version 1.0) for inter-annotator agreement and for performance in data-driven dependency parsing, observing substantial improvements in both. Most notably, the labeled attachment (LAS) accuracies for predicates, subjects and objects increase by 14.8, 5.14 and 7.35 LAS points, respectively. The new version of HOBS thus facilitates higher quality of Croatian dependency parsing, while the formalism enables more consistent manual annotation of Croatian text on the syntactic level. Both versions are publicly available for research purposes via META-SHARE.

Our future work plans include several research directions. Since three dependency treebanks of Croatian with different syntactic formalisms now exist — two versions of HOBS and the SETimes.HR treebank (Agić and Merkler, 2013) — we want to explore the prospects of combining diverse treebanks targeting improvements in dependency parsing quality along the lines of (Johansson, 2013). Following a very recent line of work in delexicalized parsing (McDonald et al., 2013), we wish to explore the impact of rich morphosyntactic tagsets and close relatedness of languages on delexicalization in parsing using the Croatian and Slovene treebanks. We also aim at enriching the treebank with semantic annotation.

Acknowledgements This work was funded by the European Union through project XLike[5] (FP7-ICT-2011-288342).

5. References

Agić, Ž. and Merkler, D. (2013). Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. *LNCS*, 8082:560–567.

Agić, Ž., Merkler, D., and Berović, D. (2013). Parsing Croatian and Serbian by Using Croatian Dependency Treebanks. In *Proc. SPMRL*, pages 22–33.

Agić, Ž. (2012). K-Best Spanning Tree Dependency Parsing With Verb Valency Lexicon Reranking. In *Proc. COLING*, pages 1–12.

Anić, V. (2003). Veliki rječnik hrvatskoga jezika.

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V., Znika, M., et al. (1995). *Hrvatska gramatika*.

Berović, D., Agić, Ž., and Tadić, M. (2012). Croatian Dependency Treebank: Recent Development and Initial Experiments. In *Proc. LREC*, pages 1902–1906.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003). The Prague Dependency Treebank. In *Treebanks*, pages 103–127.

Bohnet, B., Nivre, J., Boguslavsky, I., Ginter, R. F. F., and Hajič, J. (2013). Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *TACL*, 1:415–428.

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proc. COLING*, pages 89–97.

Buchholz, S. and Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proc. CoNLL*, pages 149–164.

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

Hajič, J., Panevová, J., Buránová, E., Urešová, Z., Bémová, A., Štepánek, J., Pajas, P., and Kárnık, J. (2001). A Manual for Analytic Layer Tagging of the Prague Dependency Treebank.

Johansson, R. (2013). Training Parsers on Incompatible Treebanks. In *Proc. NAACL-HLT*, pages 127–137.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proc. HLT-EMNLP*, pages 523–530.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proc. ACL*, pages 92–97.

Merkler, D., Agić, Ž., and Agić, A. (2013). Babel Treebank of Public Messages in Croatian. *Procedia — Social and Behavioral Sciences*, 95:490–497.

Mikelić Preradović, N., Boras, D., and Kišiček, S. (2009). CROVALLEX: Croatian Verb Valence Lexicon. In *Proc. ITI*, pages 533–538.

Mille, S., Burga, A., Ferraro, G., and Wanner, L. (2012). How Does the Granularity of an Annotation Scheme Influence Dependency Parsing Performance? In *Proc. COLING*, pages 839–852.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007a). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proc. CoNLL Shared Task Session of EMNLP-CoNLL*, pages 915–932.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007b). MaltParser: A Language-independent System for Data-driven Dependency Parsing. *Natural Language Engineering*, 13(2):95–135.

Pranjković, I. (1986). *Koji* i *što*. *Jezik*, 34:10–16.

Raguž, D. (1997). *Praktična hrvatska gramatika*.

Silić, J., Pranjković, I., and Požgaj-Hadži, V. (2005). *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*.

Šonje, J. (2000). *Rječnik hrvatskoga jezika*.

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

[5] http://www.xlike.org/