The SETIMES.HR Linguistically Annotated Corpus of Croatian

Željko Agić,[⋆] Nikola Ljubešić[†]
[⋆]Linguistics Department, University of Potsdam, Germany
[†]Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
zagic@uni-potsdam.de, nljubesi@ffzg.hr

Abstract

We present SETIMES.HR— the first linguistically annotated corpus of Croatian that is freely available for all purposes. The corpus is built on top of the SETIMES parallel corpus of nine Southeast European languages and English. It is manually annotated for lemmas, morphosyntactic tags, named entities and dependency syntax. We couple the corpus with domain-sensitive test sets for Croatian and Serbian to support direct model transfer evaluation between these closely related languages. We build and evaluate statistical models for lemmatization, morphosyntactic tagging, named entity recognition and dependency parsing on top of SETIMES.HR and the test sets, providing the state of the art in all the tasks. We make all resources presented in the paper freely available under a very permissive licensing scheme.

Keywords: dependency treebank, Croatian language, free availability

1. Introduction

There is a very clear distinction in the field of natural language processing today between resource-rich and resource-poor languages in terms of availability of language technologies, with a large majority of world languages being classified as under-resourced (Bender, 2013). Virtually all languages of Southeast Europe are under-resourced to a certain level, including Croatian (Tadić et al., 2012). As the development of language technologies for a certain language requires substantial and diverse resources over extended periods of time — while at the same time the natural language processing community thrives on developing and testing generic solutions across languages and language groups — free availability of basic resources plays an important role in developing technologies for an under-resourced language. While a number of basic language resources and tools does exist for Croatian, a large majority of them is either available in limited form and through restrictive licensing, or not publicly available at all . A detailed account on this topic is given by Tadić et al. (2012), including a classification of basic resources for Croatian on basis of availability and maturity. Furthermore, we argue for the importance of free culture licensing[1] in supporting rapid development of language technologies for under-resourced languages with inherently small research communities, such as Croatian.

In this contribution, we present the SETIMES.HR linguistically annotated corpus of Croatian text. It is the first and — to this point and to the best of our knowledge — the only basic language resource for Croatian that is distributed under free culture licensing or, more specifically, the CC BY-SA 3.0 license.[2] It is based on contemporary Croatian newspaper text from web sources and accompanied by a number of freely available models for standard tools to facilitate the basic tasks in natural language processing. These currently include: sentence boundary detection and tokenization, lemmatization and part-of-speech tagging, named entity recognition and syntactic dependency parsing. With the first version of the corpus already freely downloadable,[3] there is ongoing further work on SETIMES.HR through which we seek to enable more advanced language technologies for Croatian to the prospective users.

The remainder of the text is a more detailed account on creating the corpus and testing it in basic natural language processing tasks. First, we present the process of its linguistic annotation, from collecting and preprocessing the text to manual morphological and syntactic annotation and named entity tagging. Second, we utilize the corpus to build and test statistical models of standard tools for basic natural language processing tasks: lemmatization and part-of-speech tagging, named entity tagging and syntactic parsing. Finally, we conclude by sketching the possible future research directions involving the resource.

2. The SETIMES.HR corpus

SETIMES.HR is based on text from the latest version of the freely available SETIMES parallel corpus.[4] It is a large parallel corpus of news in nine languages of South East Europe and English. The text constituting the SETIMES corpus is automatically processed for sentence boundaries and sentence-aligned between all ten languages. For SETIMES.HR, we chose a sequential subset of 8,000 Croatian sentences from the intersection of all SETIMES languages. We manually verified the sentence splitting and conducted automatic tokenization with subsequent manual verification

[1] http://freedomdefined.org/Licenses

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

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

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

Dataset	Tokens	Types	Lemmas	MSDs
SETIMES.HR	89,129	18,007	9,045	663
news.test.hr	2,297	1,238	992	233
news.test.sr	2,320	1,221	982	234
wiki.test.hr	1,878	995	802	192
wiki.test.sr	1,947	1,018	797	195

Table 1: Basic statistics for SETIMES.HR and the test sets

in the later annotation stages. All 8,000 selected sentences entered the process of manual annotation for named entities, while the first 4,000 of these were manually lemmatized, morphosyntactically tagged and annotated for syntactic dependencies. Together with this text, we also compiled a small collection of Croatian and Serbian newspaper text and Wikipedia articles — 100 sentences each, amounting to a total of 400 sentences when combined — to serve as test sets in this and future experiments. We chose texts from different domains (news and Wikipedia) to test the effects of domain change on the statistical models. Additionally we chose texts from two different, but closely related languages (Croatian and Serbian) to test the possibility of direct model transfer between the two languages. In terms of our introductory discussion on approaches to tackling the issues of under-resourced languages, and keeping in mind that Serbian is also an under-resourced language (Vitas et al., 2012), we make a point of testing all SETIMES.HR-derived models (except NER models which is considered future work) on both Croatian and Serbian test sets. The basic statistics for the corpus and the test sets are given in Table 1. The datasets are all encoded in the style of the ConLL 2006 and 2007 shared tasks in multilingual data-driven dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007a).

The following three subsections describe the three layers of annotation — morphology, named entities and dependency syntax in SETIMES.HR and the accompanying test sets.

2.1. Morphosyntactic annotation

The previously sentence-split and tokenized SETIMES.HR text was annotated for base word forms (lemmas) and morphosyntactic tags (MSD tags) by a semi-automatic procedure, which was in part facilitated by the Croatian Morphological Lexicon (Tadić, 2006).[5] For each word form at unigram level, the lemmatization server provides a number of lemma-tag pairs, i.e., it performs ambiguous tagging. The tags are documented as mostly conformant, although with slight deviations, with the Multext-East version 4 (MTE v4) tagset specification for Croatian (Erjavec, 2012). Expert annotators then manually disambiguated the output, selecting the correct lemma-tag pairs in sentence contexts. In the process, they also manually corrected the few remaining errors in sentence splitting and tokenization. The manual annotation was followed by a number of post-processing steps in which we made the corpus fully conformant with MTE v4. Finally, on top of this, we implemented and documented a minor revision of MTE v4 geared towards the specifics of Croatian morphosyntax.[6] The biggest changes in this new tagset are the switch of participal adjectives and adverbs from the verbal subset to the adjectival and adverbial subsets and a general shrinking of the length of many tags in conformance with MTE v4 guidelines. This version of MTE-style annotation is available with the corpus. Some additional insight on morphosyntax in SETIMES.HR is provided byAgić et al. (2013a).

[5] http://hml.ffzg.hr/

Syntactic tag	%	Gloss
Adv	4.97	adverb
Ap	2.90	apposition
Atr	26.39	attribute
Atv	1.69	predicate complement
Aux	6.49	auxiliary
Co	3.37	coordinator
Elp	0.57	ellipsis
Obj	7.42	object
Oth	1.79	other, unclassified
Pnom	1.65	nominal predicate
Pred	9.31	predicate
Prep	9.62	preposition
Punc	13.24	punctuation
Sb	7.09	subject
Sub	3.50	subordinator

Table 2: List of syntactic tags in SETIMES.HR, their relative frequencies and glosses

2.2. Named entities

Three named entity classes are included in SETIMES.HR annotation: names of locations (LOC), organizations (ORG) and personal names (PERS), i.e., the three canonical name (ENAMEX) classes (Tjong Kim Sang and De Meulder, 2003). They were encoded in the IOB2 as it fits with the CoNLL format in a straightforward manner. A total of 14,193 named entities (NEs) — amounting to a total of 23,225 tokens constituting named entities, or 1.64 tokens per NE — were manually annotated in 8,000 sentences, with a 41-33-26 percentage split between LOC, ORG and PERS. This amounts to 1.77 named entities per sentence, which is an expectedly dense distribution for a newspaper text. More details on named entities in SETIMES.HR can be found in (Ljubešić et al., 2013).

2.3. Syntactic annotation

Syntactic annotation of SETIMES.HR was built on top of the manual morphological annotation as described in the previous section. We implement a standard shallow syntactic representation in form of dependency trees by the attachment of sentence dependents to the respective heads using link labels as syntactic functions of the dependents. Drawing from the recent experiences in Croatian dependency treebanking (Agić and Merkler, 2013; Agić et al., 2013b) and the experience of the JOS corpus of Slovene (Erjavec et al., 2010), in SETIMES.HR we depart from the ubiquitous Prague Dependency Treebank (Böhmová et al., 2003)

[6] http://nlp.ffzg.hr/data/tagging/msd-hr.html

	news.test		wiki.test	
LEM	hr	sr	hr	sr
CST	**97.78**	**95.95**	**96.59**	96.30
+ lex	97.04	95.52	96.38	**96.61**
POS				
HunPos	**97.04**	**95.47**	94.25	**96.46**
+ lex	96.60	95.09	**94.62**	95.58
MSD				
HunPos	**87.11**	**85.00**	**80.83**	**82.74**
+ lex	84.81	81.59	78.49	79.20

Table 3: Overall lemmatization and tagging accuracy for CST and HunPos with and without the additional inflectional lexicon by Apertium

syntactic annotation guidelines and use a simplistic 15-tag syntactic tagset described by Merkler et al. (2013). The tag descriptions and their relative frequencies in SETIMES.HR are given in Table 2. The tagset itself is aimed at improved parsing accuracy in downstream natural language processing tasks for which an expressive syntactic formalism is not required, while it is still attainable through the rich MTE-based morphosyntactic tags. The annotation process was manual, and it was conducted using DgAnnotator.[7] A detailed comparison of the syntactic layer of SETIMES.HR with the PDT-based Croatian Dependency Treebank (Tadić, 2007) in terms of supporting inter-annotator agreement and statistical dependency parsing is given by Agić and Merkler (2013), establishing a strong preference for SETIMES.HR, while Agić et al. (2013b) compare the two treebanks in the task of direct model transfer for dependency parsing of Serbian, maintaining the same preference.

3. Experiments

In this section, we present the results of experiments in which we used the SETIMES.HR corpus to train statistical models for lemmatization, morphosyntactic tagging, named entity recognition and syntactic dependency parsing using standard natural language processing tools. In all experiments, we used SETIMES.HR to derive the models and the four test sets to evaluate the models, the exception being named entity recognition, in which we test only on Croatian data.

3.1. Lemmatization and tagging

In building statistical models for lemmatization and morphosyntactic tagging, i.e., POS and full MSD tagging, we use the publicly available CST (Ingason et al., 2008) and HunPos (Halácsy et al., 2007) systems. The choice is based on previous research (Agić et al., 2013a), in which these tools were selected as top-performers in an evaluation involving a number of freely available lemmatizers and taggers. We also utilize the Apertium inflectional lexicon of Croatian (Peradin and Tyers, 2012) as both tools implement the option of adding such external information sources for improved annotation.

The results are given in Table 3. Lemmatization reaches an average score of approximately 97%, which is the state of the art for both Croatian and Serbian text. POS tagging peaks at approximately 96% and MSD tagging at 87%, which also presents the state of the art for the given languages. Language and domain shifts may be observed in the scores, with the domain influence taking precedence over the language influence, strongly supporting the option of direct model transfer. As a side note, we observe no contribution in introducing the Apertium inflectional lexicon as an additional source of information, in turn making the models more portable and independent.

[7] http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/

	news.test		wiki.test	
	hr	sr	hr	sr
LAS	76.72	75.45	71.91	72.44
UAS	81.65	80.60	80.03	80.67

Table 4: Overall dependency parsing accuracy

3.2. Named entity recognition

Named entity recognition using SETIMES.HR is facilitated by the Stanford Named Entity Recognizer (Finkel et al., 2005). We use only token-level features without additional linguistic annotation and pair SETIMES.HR with a distributional similarity clustering model by (Clark, 2003) over a 10 Mw sub-corpus of the Croatian web corpus (Ljubešić and Erjavec, 2011). We reach an overall F_1 score of approximately 90% which is maintained over the three ENAMEX categories. A detailed insight into the experiments with Croatian and Slovene named entity recognition is provided by Ljubešić et al. (2013), where SETIMES.HR is combined with other available datasets for improved NE tagging.

3.3. Dependency parsing

For dependency parsing, we use the non-projective parsing algorithm of the graph-based parser generator MSTParser (McDonald et al., 2005) as preferred for highly non-projective languages (Agić, 2012) over standard transition-based systems such as MaltParser (Nivre et al., 2007b). We observe the standard dependency parsing accuracy metrics of labeled (LAS) and unlabeled attachment (UAS).

The scores are provided in Table 4. Once again, we observe that the domain shift causes a more substantial accuracy decrease than the language shift. The top LAS and UAS scores present the state of the art in Croatian parsing and, together with (Agić et al., 2013b), the first documented scores for dependency parsing of Serbian. Moreover, using MSTParser in a standard — and slightly less rigid than our cross-language test framework — tenfold cross-validation experiment on SETIMES.HR, we reach a LAS score of slightly above 80 points, while using a more advanced graph-based parser of (Bohnet, 2010), we peak at approximately 83 LAS points. There is a line of research in statistical dependency parsing of Croatian that contains further details into the topic (Berović et al., 2012; Agić, 2012; Agić and Merkler, 2013; Agić et al., 2013b).

4. Conclusions and future work

In this paper, we have presented SETIMES.HR— the first and, to the best of our knowledge, the only linguistically annotated corpus of Croatian that is freely available for both research and commercial purposes. Through it, we seek to facilitate lower entry requirements in the development of more advanced language technologies for Croatian. Using the corpus, we built a number of models for standard natural language processing tools, deriving state-of-the-art systems for lemmatization, morphosyntactic tagging, named entity recognition and syntactic dependency parsing of Croatian text.

Our future work plans include both enlarging the corpus and enriching it with further layers of annotation, and also applying and benchmarking more advanced natural language processing tools to couple with upcoming distributions of the corpus.

The SETIMES.HR corpus, the test sets and all the natural language processing models documented in this paper are freely downloadable under the CC BY-SA 3.0 license.[8]

Acknowledgements The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement no. PIAP-GA-2012-324414 (project Abu-MaTran).

5. References

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

Agić, Ž., Ljubešić, N., and Merkler, D. (2013a). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proc. BSNLP*, pages 48–57.

Agić, Ž., Merkler, D., and Berović, D. (2013b). 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.

Bender, E. (2013). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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. (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.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proc. EACL*, pages 59–66.

Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *Proc. LREC*.

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

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. ACL*, pages 363–370.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An Open Source Trigram Tagger. In *Proc. ACL*, pages 209–212.

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *Proc. GoTAL*, pages 205–216.

Ljubešić, N. and Erjavec, T. (2011). hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. In *Proc. TSD*, pages 395–402.

Ljubešić, N., Stupar, M., Jurić, T., and Agić, Ž. (2013). Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):35–57.

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.

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

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.

Peradin, H. and Tyers, F. (2012). A Rule-based Machine Translation System from Serbo-Croatian to Macedonian. In *Proc. FREERBMT*, pages 55–65.

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

Tadić, M. (2006). Croatian Lemmatization Server. In *Proc. FASSBL*, pages 140–146.

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

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proc. CoNLL*, pages 142–147.

Vitas, D., Krstev, C., Obradović, I., Pavlović-Lažetić, G., and Stanojević, M. (2012). *The Serbian Language in the Digital Age*. White Paper Series. Springer.

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