# Lemmatization and Morphosyntactic Tagging of Croatian and Serbian

**Željko Agić*** **Nikola Ljubešić*** **Danijela Merkler†**
*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 nljubesi@ffzg.hr dmerkler@ffzg.hr

## Abstract

We investigate state-of-the-art statistical models for lemmatization and morphosyntactic tagging of Croatian and Serbian. The models stem from a new manually annotated SETIMES.HR corpus of Croatian, based on the SETimes parallel corpus. We train models on Croatian text and evaluate them on samples of Croatian and Serbian from the SETimes corpus and the two Wikipedias. Lemmatization accuracy for the two languages reaches 97.87% and 96.30%, while full morphosyntactic tagging accuracy using a 600-tag tagset peaks at 87.72% and 85.56%, respectively. Part of speech tagging accuracies reach 97.13% and 96.46%. Results indicate that more complex methods of Croatian-to-Serbian annotation projection are not required on such dataset sizes for these particular tasks. The SETIMES.HR corpus, its resulting models and test sets are all made freely available.

## 1 Introduction

Part of speech tagging (POS tagging) is an natural language processing task in which words are annotated with the corresponding grammatical categories – parts of speech: verb, noun, adjective, pronoun, etc. – in a given context. It is also frequently called morphosyntactic tagging (MSD tagging, i.e., tagging with morphosyntactic descriptions), especially when addressing highly inflected languages, for which the tagging process often includes assigning additional subcategories to words, such as gender and case for nouns or tense and person for verbs. POS/MSD tagging is a well-known task and an important preprocessing step in natural language processing. It is often preceded or followed by lemmatization – the process of mapping inflected word forms to corresponding base forms or lemmas. State of the art in POS/MSD tagging and lemmatization across languages is generally achieved – both in terms of per token accuracy and speed and robustness – by statistical methods, which involve training annotation models on manually annotated corpora.

In this paper, we investigate the possibility of utilizing statistical models trained on corpora of Croatian in lemmatization and MSD tagging of Croatian and Serbian. We present a new manually annotated corpus of Croatian – the SETIMES.HR corpus. We test a number of lemmatizers and MSD taggers on Croatian and Serbian test sets from two different domains and consider options of annotation transfer between the two languages. We also outline a first version of the Multext East v5 tagset and three usable reductions of this tagset. Special emphasis is given to rapid resource development and public availability of our research. Thus, the SETIMES.HR corpus, the test sets and the best lemmatization and MSD tagging models are made publicly available.[1] In the following section, we discuss related work on lemmatization and tagging of Croatian and Serbian. We then present the SETIMES.HR corpus and the test sets, selected lemmatizers and morphosyntactic taggers and the experimental method. Finally we provide a discussion of the evaluation results and indicate future work directions.

## 2 Related work

The task of tagging English sentences with parts of speech is generally considered a closed issue. This is due to the fact that, over the course of the past 11 years, from (Brants, 2000) to (Søgaard, 2011), the current state of the art in tagging English has improved by 1.04 – to 97.50% in terms of per token accuracy. This is, however, not the case for languages with richer morphology and free sentence

[1] http://nlp.ffzg.hr/resources/models/



*Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57,
Sofia, Bulgaria, 8-9 August 2013. 

word order, such as Croatian and Serbian.

Current state of the art for statistical MSD tagging of Croatian is reported at 86.05% (Agić et al., 2008). It involves a hidden Markov model trigram tagger CroTag, trained on the Croatia Weekly 100 thousand wordform (100 kw) subcorpus of Croatian newspaper text from Croatian National Corpus (Tadić, 2009), manually MSD-tagged and lemmatized using the Multext East v3 tagset (MTE v3) (Erjavec, 2004) and Croatian Lemmatization Server (Tadić, 2005) for guided annotation. The tagger is not publicly available. Just recently, the Croatia Weekly corpus has been made publicly available through META-SHARE.[2] Another line of research reports on a prototype constraint grammar tagger for Croatian (Peradin and Šnajder, 2012), which scores at 86.36% using a MTE-based tagset. This tagger is also not publicly available as it is in prototype stage and it currently does not analyze out-of-vocabulary word forms. The top score for lemmatizing Croatian text is reported at 96.96% by combining CroTag and Croatian Morphological Lexicon (Agić et al., 2009). The lemmatizer is not publicly available.

Lemmatization and tagging of Serbian text was recently addressed in (Gesmundo and Samardžić, 2012a; Gesmundo and Samardžić, 2012b). It involves BTagger, a combined bidirectional tagger-lemmatizer tool which implements a lemmatization-as-tagging paradigm. Models are trained on the Serbian Multext East 1984 corpus, they are publicly available[3] under a permissive license, reaching overall accuracies of 97.72% for lemmatization and 86.65% for MSD tagging. It should be noted, however, that BTagger evaluation in terms of spatial and temporal complexity was not documented and that the results provided for Serbian are obtained on specific in-domain data, i.e., a corpus of fiction and are thus not directly comparable to, e.g., results for Croatian on the Croatia Weekly newspaper corpus.

Other lines of research in Serbian lemmatization and tagging exists. Delić et al. (2009) deals with transformation-based tagging of Serbian text, but it does not provide state-of-the-art results or freely available resources. Rule-based approaches to processing Serbian using NooJ [4] and similar linguistic development environments have been thoroughly explored (Vitas et al., 2003). Several resources relevant for Serbian lemmatization and tagging are provided to the public. The Serbian version of Jules Verne 60 kw manually lemmatized and MTE-tagged corpus implements a small deviation from MTE v4 and deals with specific fictional closed-vocabulary data. SrpLemKor is a 3.7 Mw corpus of Serbian newspaper text, automatically lemmatized and POS-tagged using TreeTagger (Schmid, 1995) with a tagset of 16 POS tags. A morphological dictionary of 85 thousand Serbian lemmas with sligtly deviated MTE v4 tagset is available through NooJ. Public availability of these resources is enabled through META-SHARE, with somewhat more restrictive licensing that involves non-commercial use in all cases and for some of them it also imposes no redistribution.

Related work on lemmatizer and tagger comparison exists for many languages. Restraining the search to closely related Slavic languages, extensive work in this domain has been done for Bulgarian (Georgiev et al., 2012), Czech (Spoustová et al., 2007) and Slovene (Erjavec and Džeroski, 2004; Rupnik et al., 2008). For Croatian, preliminary work on tagger evaluation for tagger voting has been conducted (Agić et al., 2010).

## 3 SETIMES.HR corpus

SETIMES.HR is a new manually lemmatized and MSD-tagged corpus of Croatian. It is built on top of the SETimes parallel newspaper corpus involving 10 languages from the SEE region,[5] Croatian and Serbian included. This initial dataset selection was deliberate in terms of enabling us with possibility of cross-lingual annotation projection and other cross-lingual experiments. SETIMES.HR was annotated by experts using the Croatian Lemmatization Server (HML)[6] (Tadić, 2005) to facilitate the process. We made a number of changes to the initial annotation provided by human annotators. Namely, HML provides MSD tags using an undocumented alteration of the initial MTE tagset, which we corrected to conform entirely to the MTE v4 standard (Erjavec, 2012). Also, for certain lemmas HML provides lemmatization with morphosemantic cues encoded by lemma numbering – e.g. *biti1* (en. *to be*) and *biti2* (en. *to beat*) – which we omitted as they are used only in the process of generating the morphological lexicon (Tadić and Fulgosi, 2003)

[2] http://metashare.elda.org/

[3] https://github.com/agesmundo/BTagger

[4] http://www.nooj4nlp.net/

[5] http://www.nljubesic.net/resources/corpora/setimes/

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

| Corpus | Sent's | Tokens | Types | Lemmas |
|---|---|---|---|---|
| SETIMES.HR | 4 016 | 89 785 | 18 089 | 8 930 |
| set.test.hr | 100 | 2 297 | 1 270 | 991 |
| set.test.sr | 100 | 2 320 | 1 251 | 981 |
| wiki.test.hr | 100 | 1 887 | 1 027 | 802 |
| wiki.test.sr | 100 | 1 953 | 1 055 | 795 |

Table 1: Stats for SETIMES.HR and test sets

| | | set.test | | wiki.test | |
|---|---|---|---|---|---|
| Tagset | SETIMES.HR | hr | sr | hr | sr |
| MTE v4 | 660 | 235 | 236 | 188 | 192 |
| MTE v5 | 663 | 233 | 234 | 192 | 195 |
| MTE v5r1 | 618 | 213 | 216 | 176 | 180 |
| MTE v5r2 | 634 | 216 | 217 | 178 | 181 |
| MTE v5r3 | 589 | 196 | 199 | 162 | 166 |

Table 2: Tagset variation in tag counts

and are thus not required for purposes of lemmatization and MSD tagging. We make the resulting 90 kw SETIMES.HR corpus, along with the four test sets, publicly available under the CC-BY-SA-3.0 license.[7] Corpus stats are given in Table 1.

For purposes of this experiment, we propose an alteration of the baseline MTE v4 tagset in form of a first version for the MTE v5 standard.[8] The biggest changes in the new version are participal adjectives and adverbs moving from the verbal subset – which was very complex in v4 – to the adjectival and adverbial subsets. Additionally, acronyms are moved from the abbreviation subset to the noun subset. A general shrinking of the length of many tags was performed as well because from v4 onwards the MTE standard does not require one tagset for all languages in the standard. We also suggest three reductions of the suggested MTE v5 tagset:

1. without adjective definiteness (v5r1),
2. without common (Nc) vs. proper (Np) distinction for nouns (v5r2) and
3. without both (v5r3).

Adjectival definiteness is a category which is easy to implement in a morphological lexicon, but is very hard to distinguish in context as many of its variants are homographs. We question the distinction between common and proper nouns as well since they are contextually very hard to discriminate. On the other hand, some foreign proper nouns are inflected by specific paradigms and suffix tries used on unknown words could profit from this distinction. Stats for the MTE v5 and the reduced tagset versions in comparison with the baseline MTE v4 tagset version of SETIMES.HR are given in Table 2. They reflect the design choices we made: MTE v5 has a comparable amount of tags as MTE v4, gaining additional tags in the adjective subset, but losing tags in the verb and abbreviation subsets, while the reductions subsequently lower the overall MSD tag count.

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

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

## 4 Experiment setup

In this section, we define specific experiment goals and the experiment design. We also present the datasets and tools used in the experiment.

### 4.1 Objectives

The principal goal of this experiment is to provide prospective users with freely available – downloadable, retrainable and usable, both for research purposes and for commercial use – state-of-the-art lemmatization and tagging modules for Croatian and Serbian. An additional goal of our experiment is to inspect lemmatization and tagging tools available under permissive licenses and give an overview regarding their accuracy and time complexity when used on languages of morphological complexity such as Croatian and Serbian.

Regarding the previously discussed constraints on existing corpora and tools for Croatian and Serbian tagging and lemmatization, our objective implies exclusive usage of the SETIMES.HR corpus in the experiment.[9] Since SETIMES.HR is part of the SETimes parallel corpus which, among other languages, includes both Croatian and Serbian, manually annotated SETIMES.HR text has a freely available Serbian equivalent. Our first course of action was thus to train a number of taggers and lemmatizers on SETIMES.HR and test it on Croatian and Serbian held out text to verify state-of-the-art accuracy on Croatian text and to observe whether the expected decline in accuracy on Serbian text is substantial or not.

In case of substantial decrease in accuracy for lemmatizing and tagging Serbian using Croatian models, we designed multiple schemes for projecting annotation from SETIMES.HR to its Serbian

[9] Considering corpora of Croatian and Serbian stated in related work, we chose not to use non-MTE resources and corpora of fiction as an experiment basis. Importance of encoding the full set of morphological features from the MTE tagset is illustrated by its benefits for dependency parsing of Croatian (Agić and Merkler, 2013).

equivalent from the SETimes parallel corpus. The general directions for identifying the bitext subset for annotation projection were using parallel sentences which have the highest longest common subsequence or using statistical machine translation to produce Serbian sentences with minimum difference to the Croatian counterpart. Projecting tags on a bitext of high similarity would include heuristics of annotating the variation with the same morphosyntactic category if the variation was one token long or annotating it with the existing model for tagging if the variation was longer than that. Lemmatization of the single-token variation would be reapplied if the token ending in both languages was identical while other cases would be annotated with the existing lemmatization model.

### 4.2 Experiment workflow

We do four batches of experiments:

1. to identify the best available tool and underlying paradigm for lemmatization and tagging of both languages by observing overall accuracy and execution time,
2. to establish the need for annotation projection from Croatian SETIMES.HR corpus to its Serbian counterpart,
3. to select the best of the proposed MTE-based tagsets for both tasks and
4. to provide in-depth evaluation of the selected top-performing lemmatizer and tagger on both languages by using the top-performing tagset.

In the first experiment batch, we test the tools only on Croatian data from SETimes. The second batch establishes the need for – or needlessness of – annotation projection for improved processing of Serbian text by testing the tools selected in the first batch on both languages. The in-depth evaluation of the third and fourth experiment batch includes, for both languages and all test sets, observing the influence of tagset selection to overall accuracy and investigating tool performance in more detail. We measure precision, recall and $F_1$ scores for selected parts of speech and inspect lemmatization and tagging confusion matrices for detailed analysis and possible prediction of tool operation in real-world language processing environments.

We aim for the experiment to serve as underlying documentation for enabling prospective users in implementing more complex natural language processing systems for Croatian and Serbian by using these resources. Additionally, the overview of the usability of tools available is informative for researchers developing basic language technologies for other languages. We test statistical significance of observed differences in our results by using the approximate randomization test.

### 4.3 Datasets

All models are trained on SETIMES.HR. To at least partially avoid the possible pitfall of exclusive in-domain testing, we define two test sets for each language. The first test set consists of 100 Croatian-Serbian parallel sentence pairs taken by random sampling from the relative complement of the SETimes parallel corpus and SETIMES.HR. The second test set is taken from the Croatian and Serbian Wikipedia by manually selecting 20 matching Wikipedia articles and manually extracting 100 approximate sentence pairs. We chose manual over random sampling from Wikipedia to account for the fact that a certain number of articles is virtually identical between the two Wikipedias due to language similarity and mutual copying between Wikipedia users. All four test sets were manually annotated using the same procedure that was used for SETIMES.HR. The stats are given in Table 1. In addition, we have verified the difference between language test sets by measuring lexical coverage using HML as a high-coverage morphological lexicon of Croatian. For the Croatian SETimes and Wikipedia samples, we detected 5.2% and 3.9% out-of-vocabulary word forms and 11.40% and 8.86% were observed for the corresponding Serbian samples, supporting well-foundedness of the test sets in terms of maintaining the differences between the two languages.

### 4.4 Lemmatizers and taggers

As lemmatizers and taggers with permissive licensing schemes and documented cross-lingual state-of-the-art performance have become largely available, we chose not to implement our own but to obtain a set of tools and test them using our data, i.e., train them on the SETIMES.HR corpus and test them on Croatian and Serbian SETimes and Wikipedia test samples. We selected the tools on the basis of availability and underlying stochastic paradigms as to identify the best tools and best paradigms.

We tested hidden Markov model trigram taggers HunPos[10] (Halácsy et al., 2007) and lemmatization-capable PurePos[11] (Orosz and Novák, 2012),

[10] https://code.google.com/p/hunpos/

[11] https://github.com/ppke-nlpg/purepos

| Tool | Lem. | MSD | Train (sec) | Test (sec) |
|---|---|---|---|---|
| BTagger | 96.22 | 86.63 | 24 864.47 | 87.01 |
| CST | **97.78** | – | **1.80** | **0.03** |
| + lex | 97.04 | – | 1.87 | 0.12 |
| HunPos | – | **87.11** | **1.10** | **0.11** |
| + lex | – | 84.81 | 10.79 | 0.45 |
| PurePos | 74.40 | 86.63 | 5.49 | 4.42 |
| SVMTool | – | 84.99 | 1 897.08 | 3.28 |
| TreeTagger | 90.51 | 85.07 | 7.49 | 0.19 |
| + lex | 94.12 | 87.01 | 17.48 | 0.31 |

Table 3: Preliminary evaluation

lemmatization-capable decision-tree-based TreeTagger[12] (Schmid, 1995), support vector machine tagger SVMTool[13] (Giménez and Màrquez, 2004) and CST's[14] data-driven rule-based lemmatizer (Ingason et al., 2008). Keeping in mind the previously mentioned state-of-the-art scores on Serbian 1984 corpus and statistical lemmatization capability, we also tested BTagger (Gesmundo and Samardžić, 2012a; Gesmundo and Samardžić, 2012b). Since some lemmatizers and taggers are capable of using an external morphological lexicon, we used a MTE v5r1 version of Apertium's lexicon of Croatian[15] (Peradin and Tyers, 2012) where applicable.[16] All tools are well-documented and successfully applied across languages, as indicated in related work.

## 5 Results and discussion

A discussion of the experiment results follows in the next four subsections. Each subsection represents one batch of experiments. First we select the best lemmatizer and tagger, next we check for a need of annotation projection to the Serbian corpus, then the best MTE-based tagset using the best tool combination. Finally we provide a more detailed insight into the results of the top-performing pair of selected tools and tagset.

### 5.1 Tool selection

Results of the first experimental batch, consisting of testing the selected set of lemmatizers and taggers on the MTE v5r1 version of Croatian SETimes test set, are given in Table 3. In terms of lemmatization and tagging accuracy as well as processing speed in both training and testing, the top performing tools are CST lemmatizer and HunPos tagger. Thus, we chose these two for further investigation in the following batches of experiments. It should be noted that, even though its performance is comparable to the one of CST and HunPos, BTagger was not chosen for the other batches primarily because of its temporal complexity, as it is orders of magnitude higher than for the selected tools. Given that lemmatization and tagging are considered prerequisites for further processing of text tata, the data itself often being fed to these modules in large quantities (e.g., web corpora), we insist on the significance of temporal complexity in tool selection. The other results are comparable with previous research in tagging Croatian. Where applicable, we tried assisting the tools by providing Apertium's lexicon as an optional input for improved lemmatization and tagging. Only TreeTagger lemmatization and tagging benefited from lexicon inclusion. However, it should be noted that TreeTagger implements a very simple approach to lemmatization, as it only performs dictionary matching and does not lemmatize unknown words. Inclusion of a larger lexicon such as HML might be more beneficial for all the tools.

| | set.test | | wiki.test | |
|---|---|---|---|---|
| POS | hr | sr | hr | sr |
| 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 4: Overall tagging accuracy with and without the inflectional lexicon

| | set.test | | wiki.test | |
|---|---|---|---|---|
| Model | hr | sr | hr | sr |
| CST | **97.78** | **95.95** | **96.59** | 96.30 |
| + lex | 97.04 | 95.52 | 96.38 | **96.61** |

Table 5: Overall lemmatization accuracy with and without the inflectional lexicon

### 5.2 Annotation projection

HunPos tagging accuracy on all Croatian and Serbian test sets for both POS only and full MSD is given in Table 4 for the default variant and for the

[12] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[13] http://www.lsi.upc.edu/ nlp/SVMTool/

[14] http://cst.dk/online/lemmatiser/uk/

[15] http://www.apertium.org/

[16] As with already existing Croatian annotated corpora, HML is not fully MTE compliant. For future work, we might utilize a compliant version in our experiment and resulting models, being that its coverage is generally greater than the one of Apertium's lexicon due to size difference.

| Tagset | set.test | | wiki.test | |
|---|---|---|---|---|
| POS | hr | sr | hr | sr |
| MTE v4 | 96.08 | 94.61 | 93.96 | 95.85 |
| MTE v5 | 97.04 | 95.52 | **94.30** | 96.40 |
| MTE v5r1 | 97.04 | 95.47 | 94.25 | **96.46** |
| MTE v5r2 | 97.00 | **95.60** | 94.20 | 96.30 |
| MTE v5r3 | **97.13** | 95.56 | 94.09 | 96.15 |
| MSD | | | | |
| MTE v4 | 86.24 | 83.45 | 80.45 | 81.98 |
| MTE v5 | 86.77 | 84.48 | 80.46 | 82.43 |
| MTE v5r1 | 87.11 | 85.00 | 80.83 | 82.74 |
| MTE v5r2 | 87.11 | 84.96 | 81.20 | 82.38 |
| MRE v5r3 | **87.72** | **85.56** | **81.52** | **82.79** |

Table 6: HunPos POS and MSD tagging accuracy for all tagsets

| | set.test | | wiki.test | |
|---|---|---|---|---|
| Tagset | hr | sr | hr | sr |
| MTE v4 | 97.78 | 95.82 | 96.66 | 96.11 |
| MTE v5 | 97.82 | 95.86 | **96.81** | **96.30** |
| MTE v5r1 | 97.78 | 95.95 | 96.59 | **96.30** |
| MTE v5r2 | **97.87** | **95.99** | 96.75 | 96.20 |
| MTE v5r3 | 97.74 | **95.99** | 96.54 | 96.20 |

Table 7: CST lemmatization accuracy for all tagsets

one using Apertium's lexicon. These results serve as the first decision point regarding the need for Croatian-to-Serbian annotation projection, the second one being the lemmatization scores in Table 5. Here we observed an unsubstantial decrease in POS and MSD tagging between Croatian and Serbian test sets – the observed difference is, in fact, more substantial across domains than across languages. Overall, Croatian and Serbian scores differ less than 3%. Results for Serbian Wikipedia sample are even consistently better than for Croatian Wikipedia, emphasizing domain significance over language difference. The tagger does not benefit from the inclusion of the inflectional lexicon in POS tagging and it even incurs a substantial 2% to 4% penalty in MSD tagging. Since such observations were not made while including the lexicon with the TreeTagger tool – which implements the simplest form of dictionary lemmatization – we performed a small results analysis and noticed an unnaturally high percentage of categories that are as expected present in the lexicon, but very rare in the training corpus (like the vocative case) pointing to a naïve implementation of the procedure. Thus we chose not to use the lexicon in further observations. Lack of more substantial differences

| Tagsets | v5 | v5r1 | v5r2 | v5r3 |
|---|---|---|---|---|
| v4 | 0.268 | <0.05 | <0.05 | <0.01 |
| v5 | / | <0.01 | <0.05 | <0.01 |
| v5r1 | / | / | 0.877 | <0.05 |
| v5r2 | / | / | / | <0.01 |

Table 8: Statistical significance of differences in full MSD tagging between tagsets (p-values using approximate randomization)

in tagging scores between Croatian and Serbian for this specific test scenario implied no need for annotation projection.

This is further supported by overall lemmatization scores in Table 5. Even with the observed lexical differences between the languages, as we indicated in the description of the test sets by measuring lexical coverage using HML, the learned CST lemmatizer rules are more robust considering language alteration than the trigram tagging model of HunPos. Lemmatization accuracy stays in the margins of approximately $97\%\pm1\%$ for both languages. Average accuracy on Croatian is less than 2% higher than for Serbian and the domain patterns observed for tagging are also observed for lemmatization. Benefits of an inflectional lexicon for lemmatization are minor, if any, which can be followed back to the small size of the lexicon and high quality of the CST lemmatizer. On the contrary, TreeTagger's simple lemmatization does gain four points by using the lexicon, but it initially performs seven points worse than CST.

### 5.3 Tagset selection

Tables 6 and 7 show the influence of tagset design on tagging and lemmatization accuracy. They are accompanied by Table 8, i.e., results of testing statistical significance of differences between the tagsets in the task of full MSD tagging from Table 6. Statistical significance is calculated with all test sets merged into one. Differences in lemmatization accuracy are virtually non-existent regarding the tagset choice. Full MSD tagging follows the usual pattern of inverse proportionality between tagset size and overall accuracy. It should be noted that MTE v5 accuracy is not significantly higher than MTE v4 accuracy ($p = 0.268$), but we consider the new tagset to be easier to use for humans since its tags are shortened by removing placehodlers for features used in other MTE languages. Considering that only tagging accuracy using the MTE v5r3 tagset is significantly better than tagging using all

| | Croatian | | | Serbian | | |
|---|---|---|---|---|---|---|
| POS | P | R | $F_1$ | P | R | $F_1$ |
| Adj | 94.33 | 90.14 | 92.19 | 94.34 | 93.98 | 94.16 |
| | 66.80 | 63.83 | 65.28 | 66.79 | 66.54 | 66.66 |
| Adv | 84.56 | 82.73 | 83.63 | 82.57 | 73.77 | 77.92 |
| | 84.56 | 82.73 | 83.63 | 82.57 | 73.77 | 77.92 |
| Conj | 95.29 | 93.82 | 94.55 | 97.92 | 95.29 | 96.59 |
| | 94.12 | 92.66 | 93.38 | 96.89 | 94.28 | 95.57 |
| Noun | 95.70 | 96.34 | 96.02 | 95.42 | 96.59 | 96.00 |
| | 76.78 | 77.30 | 77.04 | 75.38 | 76.30 | 75.84 |
| Num | 94.57 | 97.75 | 96.13 | 96.51 | 93.26 | 94.86 |
| | 91.30 | 94.38 | 92.81 | 94.19 | 91.01 | 92.57 |
| Prep | 98.10 | 99.72 | 98.90 | 98.45 | 98.70 | 98.57 |
| | 95.93 | 97.52 | 96.72 | 94.30 | 94.55 | 94.42 |
| Pron | 95.97 | 97.54 | 96.75 | 95.78 | 97.42 | 96.59 |
| | 81.85 | 83.20 | 82.52 | 81.43 | 82.83 | 82.12 |
| Verb | 95.88 | 98.07 | 96.96 | 95.23 | 95.72 | 95.47 |
| | 93.81 | 95.96 | 94.87 | 93.36 | 93.84 | 93.60 |

Table 9: Precision (P), recall (R) and $F_1$ score for POS only (1st column) and full MSD (2nd column) on Croatian and Serbian

other suggested tagsets, we chose this tagset and tagging model for further observation of lemmatization and tagging properties in the remainder of the paper. Still, in this section, we present the results on all tagsets to serve as underlying documentation of the observed differences, mainly because of the fact that only MTE v4 is officially supported at this moment and MTE v5 is a newly-introduced prototype that displays better performance in this specific experiment.

### 5.4 In-depth analysis

In Table 9 we merge SETimes and Wikipedia test sets by language and provide POS and MSD tagging precision, recall and $F_1$ score for selected Croatian and Serbian parts of speech. In terms of POS only, the most difficult-to-tag part of speech is the adverb, followed by the adjective in both Croatian and Serbian. The other categories are consistently POS-tagged with an $F_1$ score of approximately 95% or higher. The decrease for adverbs and adjectives is somewhat more evident in precision than in recall and the POS confusion matrix for both languages, given in Table 10, shows that these two parts of speech are often mistaken for each other by the tagger. Regarding full MSD tagging using the MTE v5r3 tagset, for both languages, the lowest $F_1$ scores are observed for adjectives (approximately 66%), nouns (76%) and pronouns (82%). This is most likely due to the fact that these parts of speech have the largest tagset subsets, making it easier for the tagger to get confused.[17] Performance for other parts of speech is satisfactory, especially for verbs, keeping in mind, e.g., possible subsequent dependency parsing of the two languages. The absolute difference between POS and MSD tagging score is most substantial for adjectives (approximately 27%), indicating that certain MSD features might be triggering the decrease. This is partially supported by our tagset design investigation as dropping adjective definiteness atribute yielded substantial overall tagging accuracy increase when compared with the tagsets in which this attribute is still encoded.

In Table 10 we provide a part of speech confusion matrix for Croatian and Serbian on test sets merged by language. In Croatian test sets, the most frequent confusions are those between adjectives and nouns (28.9%), nouns and verbs (14.5%), adjectives and adverbs (11.6%) and nouns and adverbs (6.9%). In Serbian text, the tagger most frequently confuses nouns – for adjectives (21.1%), verbs (20%) and adverbs (16%). Merging the test sets by language mostly evens out the tagging differences as there is a total of 173 MSD confusions in Croatian test sets and only 3 more, i.e., 175 in the Serbian test sets.

POS scores for both languages neared the level of human error in our experiment. Keeping that in mind, upon observing the confusion instances themselves, we spotted a confusion between adjectives and nouns (e.g. names of countries (*Hrvatska* (en. *Croatia*, *Croatian*)), homographic forms (*strana* (en. *foreign*, *side*), *svet* (en. *world*, *holy*)) and confusion between adjectives and adverbs. Adverbs and prepositions are sometimes confused with nouns, especially for nouns in instrumental case (e.g. *godinama* (en. *year*, *yearly*), *tijekom* (en. *duration*, *during*)). Conjuctions are at times incorrently tagged because various words can have a conjuctional function, most frequently pronouns and adverbs: *što* (en. *what*), *kako* (en. *how*), *kada* (en. *when*). Interestingly, there is some confusion between nouns and verbs in Wikipedia test sets, while in SETimes test sets there are almost none. This confusion arises from the homographic forms – e.g. *mora* (en. *must*, *seas*) – or from nouns with

[17] There are 589 MTE v5r3 tags in SETIMES.HR. Out of these, 164 are used for tagging adjectives, 42 for nouns and 268 for pronouns, thus accounting for 80.47% of the tagset. There are also 50 verb tags.

| POS | Abbr | Adj | Adv | Conj | Noun | Num | Part | Prep | Pron | Res | Verb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Abbr | | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| Adj | 0 | | 20 | 0 | 50 | 0 | 1 | 0 | 3 | 1 | 4 |
| Adv | 0 | 10 | | 9 | 12 | 0 | 0 | 2 | 0 | 0 | 2 |
| Conj | 0 | 0 | 5 | | 2 | 0 | 5 | 5 | 7 | 0 | 0 |
| Noun | 0 | 37 | 28 | 0 | | 4 | 0 | 1 | 5 | 7 | 25 |
| Num | 2 | 4 | 0 | 0 | 2 | | 0 | 0 | 0 | 0 | 0 |
| Part | 0 | 0 | 0 | 3 | 0 | 0 | | 0 | 0 | 0 | 3 |
| Prep | 0 | 0 | 2 | 3 | 2 | 0 | 1 | | 0 | 0 | 0 |
| Pron | 0 | 2 | 1 | 9 | 3 | 0 | 1 | 0 | | 0 | 1 |
| Res | 0 | 0 | 1 | 0 | 4 | 0 | 0 | 2 | 0 | | 0 |
| Verb | 0 | 9 | 4 | 0 | 35 | 1 | 2 | 1 | 0 | 1 | |

Table 10: POS confusion matrix for Croatian (top right) and Serbian (bottom left)

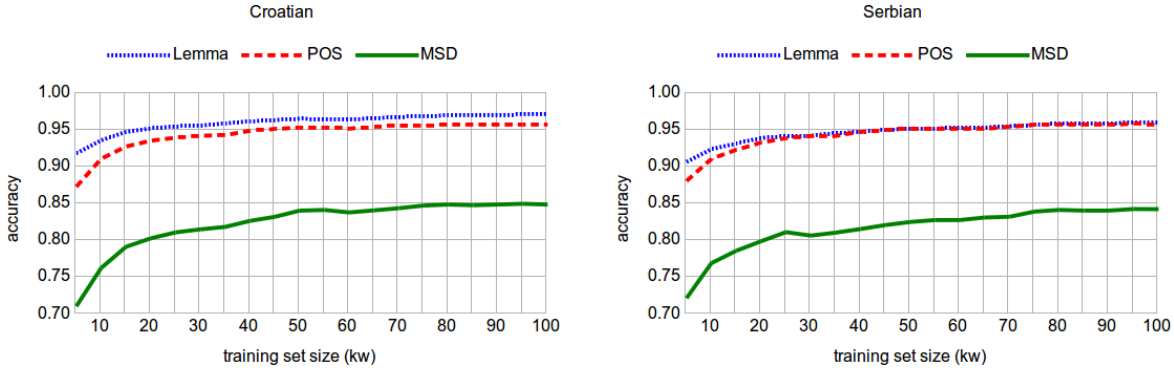


Figure 1: Learning curves for Croatian and Serbian lemmatization and tagging

suffixes *-la* and *-lo*, which are used for denoting participles in feminine and neuter gender, or with suffix *-ti*, which is also a suffix for infinitive.

Most MSD tag confusions arise from the fact that the same suffix can denote different cases in different declensions. We observed confused number and gender category (mostly in adjectives in masculine and neuter gender), but the most frequent confusion occurs for accusative forms in masculine gender, which have different suffixes when they denote animacy (suffix is the same as in the genitive case: *pobjednika* (en. *winner*), *kandidata* (en. *candidate*)) and when they denote inanimacy (suffix is the same as in the nominative case: *metak* (en. *bullet*), *bubnjar* (en. *drummer*)).

In lemmatization, as in POS tagging, errors are generally very infrequent. Some occur with adjectives, when an assigned lemma represents a definite form of an adjective, instead of an indefinitive form (and less frequently vice versa). Besides, adjectives are sometimes confused with adverbs (e.g., target lemma is *značajno* (en. *significantly*), but the lemma *značajan* (en. *significant*) is assigned, and vice versa). Other less frequent examples include cases in which the assigned lemma is not in its canonical form, but a case other than the nominative case, or when the assigned lemma is a word stem. A small number of errors also occurs due to slight differences in Croatian and Serbian wordforms, e.g., when a Serbian nominative form is not a nominative form in Croatian (*planeta* as Serbian nominative and Croatian genitive, *planet* being the Croatian nominative).

Figure 1 provides lemmatization, POS and MSD tagging learning curves for both languages on merged test sets. Apart from the slight difference in lemmatization scores in favor of Croatian, the learning curves and overall scores on merged test sets are virtually identical. The easiest task to learn is lemmatization while the most complex one is applying MSD.

## 6 Conclusions and future work

In this paper, we have addressed the issue of lemmatization and morphosyntactic tagging of two generally under-resourced languages, Croatian and Serbian. Our goal was to provide the general public with freely available language resources and state-

of-the-art models for lemmatization and tagging of these two languages in terms of accuracy, robustness and speed. We also aimed at using lemmatization and tagging as a platform for implicit comparison of the two languages in natural language processing terms, as to provide partial insight to how difficult and lossy – or, more desireably, how easy and straightforward – would it be to port linguistic resources and language processing tools from one language to another.

While developing the models, we completed a series of experiments. We used the Croatian text from the freely available SETimes parallel corpus to create a new manually lemmatized and morphosyntactically tagged corpus of Croatian – the SETIMES.HR corpus. Beside the Multext East v4 morphosyntactic tagset specification for Croatian which was used for initial corpus annotation, we designed and implemented a first version of the Multext East v5 tagset and its three reductions and applied these to SETIMES.HR. Using SETimes and Wikipedia as starting point resources, we created two gold standard test sets for each language in order to test existing state-of-the-art lemmatizers and taggers. We ran preliminary tests on a number of tools to select CST lemmatizer and HunPos tagger as tools of choice considering observed accuracy, training time and text processing time. In an in-depth evaluation of these tools, we obtained peak overall lemmatization accuracy of 97.87% and 96.30% for Croatian and Serbian and full morphosyntactic tagging accuracy of 87.72% and 85.56%, with basic part of speech tagging accuracy at 97.13% and 96.46%. In this specific test scenario and with this specific training set, we have shown the differences in results between Croatian and Serbian not to be significant enough to justify an effort in more elaborate strategy of adapting Croatian models to Serbian data – simply training the models on Croatian text from SETIMES.HR corpus and using them on Serbian text provided state-of-the-art results in lemmatization and tagging, while maintaining and even topping previously documented state of the art for Croatian.

The SETIMES.HR corpus, Croatian and Serbian test sets and top-performing lemmatization and tagging models are publicly available and freely downloadable[18] under the CC-BY-SA-3.0 license.

Our future work plans include both enlarging and enhancing SETIMES.HR. The presented learning curves show significant room for improvement by annotating additional data. The dataset already serves as a basis for the SETIMES.HR treebank of Croatian (Agić and Merkler, 2013), implementing a novel dependency syntactic formalism and enabling experiments with joint dependency parsing of Croatian and Serbian. Should dependency parsing experiments show the need for more elaborate language adaptation strategies, we will most likely implement them also on the level of lemmas and morphosyntactic tags before addressing syntactic issues. This will possibly be helped by statistical machine translation between Croatian and Serbian to enhance bitext similarity and empower projection strategies. An effort could be made to adapt existing Croatian and Serbian resources and subsequently to attempt achieving better lemmatization and tagging performance by combining these with SETIMES.HR. We will use the models presented in this paper to annotate the web corpora of Croatian and Serbian (Ljubešić and Erjavec, 2011) – hrWaC and srWaC.

[18] http://nlp.ffzg.hr/resources/models/

## Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement n° PIAP-GA-2012-324414 (project Abu-MaTran).

## References

Željko Agić and Danijela Merkler. 2013. Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. In *Text, Speech and Dialogue. Lecture Notes in Computer Science*. Springer.

Željko Agić, Marko Tadić, and Zdravko Dovedan. 2008. Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. *Informatica*, 32(4):445–451.

Željko Agić, Marko Tadić, and Zdravko Dovedan. 2009. Evaluating Full Lemmatization of Croatian Texts. In *Recent Advances in Intelligent Information Systems*, pages 175–184. Exit Warsaw.

Željko Agić, Marko Tadić, and Zdravko Dovedan. 2010. Tagger Voting Improves Morphosyntactic Tagging Accuracy on Croatian Texts. In *Proceedings of ITI*, pages 61–66.

Thorsten Brants. 2000. TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP*, pages 224–231.

Vlado Delić, Milan Sečujski, and Aleksandar Kupusinac. 2009. Transformation-Based Part-of-Speech Tagging for Serbian Language. In *Proceedings of CIMMACS*.

Tomaž Erjavec and Sašo Džeroski. 2004. Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words. *Applied Artificial Intelligence*, 18:17–41.

Tomaž Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of LREC*.

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

Georgi Georgiev, Valentin Zhikov, Kiril Simov, Petya Osenova, and Preslav Nakov. 2012. Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian. In *Proceedings of EACL*, pages 492–502.

Andrea Gesmundo and Tanja Samardžić. 2012a. Lemmatisation as a Tagging Task. In *Proceedings of ACL*.

Andrea Gesmundo and Tanja Samardžić. 2012b. Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification. In *Proceedings of LREC*.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS Tagger Generator Based on Support Vector Machines. In *Proceedings of LREC*.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: An Open Source Trigram Tagger. In *Proceedings of ACL*, pages 209–212.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *Proceedings of GoTAL*, pages 205–216.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer.

György Orosz and Attila Novák. 2012. PurePos – An Open Source Disambiguator. In *Proceedings of NLPCS*.

Hrvoje Peradin and Jan Šnajder. 2012. Towards a Constraint Grammar Based Morphological Tagger for Croatian. In *Text, Speech and Dialogue*, pages 174–182. Springer.

Hrvoje Peradin and Francis M. Tyers. 2012. A Rule-Based Machine Translation System from Serbo-Croatian to Macedonian. In *Proceedings of FREERBMT12*, pages 55–65.

Jan Rupnik, Miha Grčar, and Tomaž Erjavec. 2008. Improving Morphosyntactic Tagging of Slovene Language Through Meta-Tagging. *Informatica*, 32(4):437–444.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging With an Application to German. In *Proceedings of ACL SIGDAT Workshop*.

Anders Søgaard. 2011. Semi-Supervised Condensed Nearest Neighbor for Part-of-Speech Tagging. In *Proceedings of ACL-HLT*, pages 48–52.

Drahomíra "johanka" Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of BSNLP*, pages 67–74.

Marko Tadić and Sanja Fulgosi. 2003. Building the Croatian Morphological Lexicon. In *Proceedings of EACL 2003 Workshop on Morphological Processing of Slavic Languages*, pages 41–46.

Marko Tadić. 2005. Croatian Lemmatization Server. *Southern Journal of Linguistics*, 29(1):206–217.

Marko Tadić. 2009. New Version of the Croatian National Corpus. *After Half a Century of Slavonic Natural Language Processing*, pages 199–205.

Duško Vitas, Cvetana Krstev, Ivan Obradović, Ljubomir Popović, and Gordana 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, 1st Balkan Conference in Informatics*.