New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian

Nikola Ljubešić* **Filip Klubička*** **Željko Agić†** **Ivo-Pavao Jazbec‡**
* Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
† Center for Language Technology, University of Copenhagen
Njalsgade 140, 2300 Copenhagen S, Denmark
‡ Institute of Croatian Language and Linguistics
Ulica Republike Austrije 16, HR-10000 Zagreb, Croatia
{nljubesi,fklubick}@ffzg.hr zeljko.agic@hum.ku.dk ipjazbec@ihjj.hr

Abstract

In this paper we present newly developed inflectional lexcions and manually annotated corpora of Croatian and Serbian. We introduce *hrLex* and *srLex*—two freely available inflectional lexicons of Croatian and Serbian—and describe the process of building these lexicons, supported by supervised machine learning techniques for lemma and paradigm prediction. Furthermore, we introduce *hr500k*, a manually annotated corpus of Croatian, 500 thousand tokens in size. We showcase the three newly developed resources on the task of morphosyntactic annotation of both languages by using a recently developed CRF tagger. We achieve best results yet reported on the task for both languages, beating the HunPos baseline trained on the same datasets by a wide margin.

Keywords: inflectional lexicon, morphosyntactic annotation, Croatian, Serbian

1. Introduction

In this paper we introduce *hrLex* and *srLex*, freely available morphological lexicons of Croatian and Serbian. We describe the process of building these lexicons which includes supervised machine learning techniques for lemma and paradigm candidate ranking of non-covered words to enhance the linguists' productivity. Furthermore, we present *hr500k*, a new Croatian gold dataset manually annotated with morphosyntactic and lemma information, 500 thousand tokens in size. The dataset represents a balanced extension of the SETimes.HR dataset (Agić and Ljubešić, 2014) which consisted of newspaper articles from one source only.

We perform an intrinsic evaluation of the inflectional lexicons and an extrinsic evaluation of all three resources on the task of morphosyntactic tagging of both Croatian and Serbian. We compare the newly developed corpus with its predecessor, the SETimes.HR corpus, and measure the accuracy gain obtained by including the lexicon in the morphosyntactic tagging process. For extrinsic evaluation we use a recently developed tagger optimised on Slovene (Ljubešić and Erjavec, 2016), comparing its results to the ones obtained with the popular HunPos tagger (Halácsy et al., 2007) when trained on the same datasets.

2. Related Work

For both languages morphological lexicons were developed in the past, but with limited availability. For Croatian the Croatian Morphological Lexicon (Tadić and Fulgosi, 2003) was available for search through a web interface since 2005 (Tadić, 2005). Since 2012 this lexicon is available through Meta-Share, with a size of ca 113,000 lemmas (60% of which are proper names) in version 5.0. However, it is distributed under a non-commercial license in the form of (token, lemma, tag) triples only, and is therefore not useful for expansion or enrichment. Šnajder et al. (2008) provide another line of work on Croatian inflectional lexica, but the resulting resource is not freely available.

For Serbian the SrpMD dictionary (Krstev, 2008), 85,721 lemmas in size, is published under a non-commercial license and indexed on Meta-Share, but is not available for download.

The lexicons we present in this paper are freely downloadable, published under the GNU GPL license, organised by lexemes and paired with their inflectional paradigms, thereby enabling a wide range of applications and easy extensibility.

Similar to inflectional lexicons, the line of work in annotated corpora of Croatian is reasonably extensive, in contrast to a fairly limited amount of research carried out for Serbian (Vitas et al., 2012), especially considering syntactic annotations. By and large, however, these contributions do not result in freely available resources; for a more detailed overview, see Agić et al. (2013b). On top of providing two sizable new inflectional lexicons for the two languages, our *hr500k* corpus marks a significant new development for Croatian, and by virtue of direct transfer of tagging models, for Serbian as well. While Agić and Ljubešić (2015) document top-level results in dependency parsing, our contribution significantly improves over the previous top scores in morphosyntactic tagging for the two languages.

With these recent developments, we can safely assume that through our line of work in free-culture resources, Croatian and Serbian are leaving the realm of severely under-resourced languages.

3. Lexicon Construction

3.1. The Initial Lexicon

The morphological lexicon that acted as the starting point towards the construction of our lexicons is part of the Apertium rule-based machine translation system (Forcada et al.,

2011). This lexicon covers Bosnian, Croatian and Serbian and encodes the lexical and grammatical differences between the three languages.
This lexicon is the only freely available morphological lexicon of Bosnian, Croatian or Serbian that contains both definitions of paradigms as well as lexemes attached to these paradigms.[1]
At the time we started the constructions our two lexicons, the Apertium HBS lexicon[2] consisted of 413 paradigms from open-word classes, out of which 204 were noun paradigms, 167 were verbal and 42 adjectival. There were 10,183 lexemes in the lexicon assigned to one of the 413 paradigms. The whole lexicon had up to that point been produced manually by the members of the Apertium community.

3.2. Extending the Lexicon

We took it upon ourselves to extend this lexicon, and the benefits of this are twofold: we can use the data in our work, but we also contribute to the open-source Apertium community.
In order to identify out-of-vocabulary words (OOVs), we use the largest available corpora of Croatian and Serbian, *hrWaC* and *srWaC*, with 2 billion and 894 million tokens respectively (Ljubešić and Klubička, 2014).
We extracted the OOVs to be added based on frequency; we calculated the frequency distribution of lowercase tokens that were not already covered by the lexicon. Additionally, we implemented simple heuristics to bypass noise such as typos (via the Damerau-Levenshtein distance metric) and misspellings (in particular mistakes with diacritics and the *yat* reflexes as they are very frequent both in Croatian and Serbian).
Six linguists were hired to go through the most frequent OOVs and produce new lexicon entries in form of lemmas and their corresponding paradigms. To assist their work we used a web-based GUI presented in Figure 1 as the front end, and the predictor of lemma and inflectional paradigm for an OOV, described in Ljubešić et al. (2015), as the back end. Once presented with an OOV and pairs of lemmas and expanded paradigm candidates, the annotators could choose between one of the candidates, or flag the entry as belonging to a non-defined paradigm or a different part of speech.
We first focused on Croatian data. We had 6 rounds of paradigm annotation, and after each round we had linguists go through the flagged entries, write their paradigms or add them to the corresponding part of speech.
After having a satisfactory coverage of Croatian, we moved to Serbian and repeated the process over 2 rounds. Far less annotation rounds were needed for Serbian data due to a large lexical overlap between languages or the only difference being the *yat* reflex (e.g. Croatian *lijep*, Serbian *lep*) which was already encoded during the Croatian annotation.
After each round of annotation, the list of OOVs was regenerated, and the paradigm prediction model was retrained on the newly expanded lexicon. Thus, every upcoming round would have fresh data and would also include the paradigms that did not exist in previous rounds.

	lemmas	surface forms
hrLex	99,680	4,971,257
srLex	105,358	5,327,361

Table 1: Number of lemmas and (token, lemma, tag) triples in the hrLex and srLex lexicons

3.3. Tagset Mapping

However, even after finishing all the annotation rounds and greatly expanding the Apertium lexicon, its format is still not ideal for our purposes. One of the reasons is that it uses the Apertium tagset,[3] the usefulness of which is confined within the limits of the Apertium system. We decided it would be prudent to use a more widely accepted annotation schema, so we mapped the Apertium tagset to the MULTEXT-East Morphosyntactic Specifications, revised Version 4.[4]
Although the overall trend in the community is to switch towards the UD UPOS tagset[5], we still prefer tagging our data with the MTEv4r tagset as it is more widely accepted in the local linguistic communities.
Furthermore, we have defined a mapping from the MTEv4r tagset to the UD UPOS tagset (available together with the MTEv4r tagset definition) (Agić and Ljubešić, 2015), so moving to the other tagset can be done seamlessly.

3.4. Final Lexicons

The final lexicons are currently distributed either organised by lexeme and paired with the corresponding inflection paradigm in the Apertium (meta)dix format[6] or in form of (token, lemma, tag) triples, separately for Croatian[7] and Serbian.[8]
The number of inflectional paradigms in the Apertium lexicon has grown dramatically, and so there are now 472 noun paradigms, 568 verb paradigms and 187 adjective paradigms. The sizes of the triple-format lexicons are presented in Table 1.

3.5. Lexicon Evaluation

In order to have a better understanding of the precision of our resource, we performed a manual evaluation on a small subset of entries in the Croatian lexicon. Given that the parts of speech are far from a uniform distribution, we evaluated 1000 triples - random samples of 300 nouns, 300 verbs, 300 adjectives and 100 remaining parts of speech.

[1] http://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-hbs/

[2] HBS is the ISO 639-3 code for the macrolanguage covering the three languages in question

[3] http://wiki.apertium.org/wiki/List_of_symbols

[4] https://github.com/ffnlp/sethr/blob/master/mte4r-upos.mapping

[5] http://universaldependencies.org/u/pos/index.html

[6] http://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-hbs/apertium-hbs.hbs.metadix

[7] http://hdl.handle.net/11356/1056

[8] http://hdl.handle.net/11356/1057

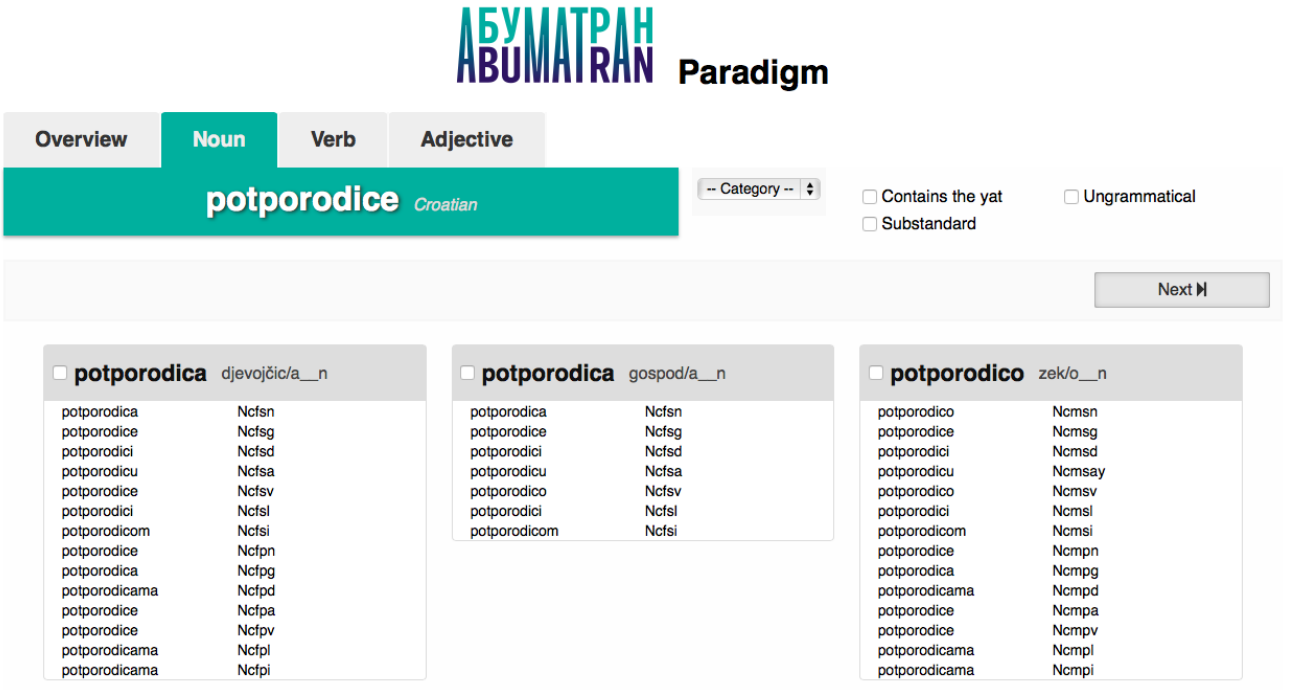

Figure 1: Example of the GUI used for extending the lexicon

	nouns	adjectives	verbs	other	average
error rate	3%	2.33%	1%	0%	1.9%

Table 2: Intrinsic evaluation by part-of-speech

The results are shown in Table 2. Most of the errors are either (1) lexical or (2) stem from an incorrect lemma-paradigm pair.

Although the average error rate of 1.9% is not ideal, it should be noted that the total time invested into the production of the lexicon is around 1500 person-hours. Given the size of the resource and its usefulness for other NLP problems (presented in section 5.), we find that the ratio of time invested on one side and precision or usefulness on the other is very acceptable.

4. Training Corpus Construction

In this section we present the *hr500k* corpus, 500 thousand tokens strong, which was manually annotated on the morphosyntax and lemma level. This resource is currently the largest training corpus of Croatian and was built in two phases.

4.1. First Phase

Our first efforts of building a training corpus for Croatian began in 2012 when a 59,212 token-strong corpus was built for the purposes of a named entity recognition task (Ljubešić et al., 2012). The data comes from four different web domains belonging to the genres of general news, ICT news and business news. These data were manually annotated during a student project where diversity of data was one of the main points.

In 2013 the SETimes.HR corpus (Agić and Ljubešić, 2014) was built as a first orchestrated effort to kickstart free culture language resources and tools for Croatian (Agić et al., 2013a). The corpus is 83,637 tokens strong and consists of newspaper articles from the multilingual and now inactive setimes.com domain.

Then in 2014, another corpus was built and tested on the task of MSD-tagging (Klubička and Ljubešić, 2014). No specific topic domain was chosen, but rather a random sample of sentences from the general web which, through our crowdsourcing efforts, were deemed as being of an acceptable linguistic standard. This dataset of 50,322 tokens was then automatically MSD-tagged, followed by employing crowdsourcing and a small team of experts to correct the annotations of tokens that were tagged differently by a tagger ensemble.

These corpora were later merged into one single corpus of approximately 190 thousand tokens in size, which was manually inspected for possible errors and inconsistencies. As for genre and register, the content of this corpus belonged mainly to news (>85%), and a little bit of the general web, which varied greatly by genre and topic, including the odd forum discussion or blog post, but mostly consisting of reports on politics, sports, religion, in addition to news and other informative articles.

4.2. Second Phase

In 2015 we set the bar to 500 thousand tokens, mostly because of our results on morphosyntactic annotation of Slovene (Ljubešić and Erjavec, 2016) which showed corpus supervision to be of much greater importance than lexicon supervision. Thus, the second phase of corpus construction consisted of manually selecting 320k tokens of suitable documents from the hrWaC web corpus and having experts do correction of automated morphosyntactic annotation learned from the 190k-sized corpus.

However, this time around we wanted the corpus to in-

	articles	blogs	forums	other
token ratio	40%	30%	20%	10%
tokens	119,745	93,335	65,941	33,290

Table 3: Tokens per web genre in the extension yielding the hr500k corpus

	articles	blogs	forums	other
token ratio	57.63%	20.6%	14.64%	7.13%
tokens	286,404	102,314	72,814	35,457

Table 4: Tokens per web genre in the hr500k corpus

clude a more varied, yet representative sample of the Croatian language; one that escapes the confines of a particular genre, topic or register, and includes many different examples of linguistic expression that can be found on the web. With accordance to that, we divided the 320k token sample into 4 sections according to web genre, in the ratios shown in Table 3.

That way, we covered the registers used in different kinds of genres – articles, blogs, forums, reviews and advertisements – while at the same time covering a wide range of topics that were inadequately or not at all covered in the initial 190k corpus (which was mainly news articles). The web domains included cover topics ranging from medicine, education and technology, through music, sports and religion, all the way to listings, literature and political activism. We took special care to include any user comments on articles and blogs, so that, coupled with forum discussions, the corpus would also include a nice sample of the language used in direct communication among internet users. Such meticulous selection results in considerable variety among documents, but at the same time the sample is still quite representative of the Croatian web, as documents were selected exclusively from a list of top 200 most frequent domains, i.e., the ones that the most documents in the hrWaC corpus come from.

An approximation of the distribution of web genres in the final hr500k corpus created by merging all the hitherto described corpora is presented in table 4. An overview of the topic domains that enriched the corpus in the second phase of construction is presented in table 5 and is based on the general topic of the web domains the sentences come from, while an approximation of topic domain distribution in the final 500k corpus is presented in table 6. Compared to the approximate >85% of general news articles that comprise the initial 190k corpus, this is a vast improvement in terms of data diversity.

topic	token ratio	topic	token ratio
general	35.01%	business	4.41%
music	13.55%	listings	3.88%
medicine	12.26%	religion	3.81%
tech	7.93%	sports	3.59%
lifestyle	7.38%	culture	2.36%
education	5.80%		

Table 5: Topic domain distribution in the 320k extension

topic	token ratio	topic	token ratio
general	51.89%	education	3.61%
music	8.43%	religion	2.87%
medicine	7.63%	sports	2.74%
business	6.93%	listings	2.42%
tech	6.92%	culture	1.97%
lifestyle	4.59%		

Table 6: Topic domain distribution in the hr500k corpus

5. Using the Resources for Morphosyntactic Annotation

In this section we present our extrinsic evaluation of the resources presented in the two previous sections. We perform the evaluation on the task of morphosyntactic tagging.

To achieve high-quality tagging of South Slavic languages we have recently developed a tagger (Ljubešić and Erjavec, 2016) from scratch as most of the available taggers have some shortcomings.

A very popular tagger, especially for inflectionally rich languages is HunPos, which was actually our tagger of choice until recently. However, HunPos is based on the HMM-based TnT tagger, lacking therefore the possibility of adding additional features. With the significant increase of available computational power, we argue for using more complex algorithms like conditional random fields that enable a richer knowledge representation. In this paper we use the HunPos tagger as a baseline.

Another possible choice nowadays is MorphoDita (Straková et al., 2014). This tagger performs very well on both Czech and English, but has the shortcoming that unknown words are handled by a separate module which is not documented, making adding new languages impossible as long as tagging of unknowns is a requirement.

A final argument for developing our own tagger was that in the near future we plan to develop taggers of non-standard language for South Slavic languages which will require adding additional features for maximising tagging accuracy.

5.1. Tagger Description

Our tagger is based on the CRF implementation CRFsuite[9] (Okazaki, 2007). We perform feature extraction both from the text to be trained on / tagged and the available lexicons. For making the lexicons space-efficient, we compiled them in form of tries using the python marisa-trie wrapper[10].

The feature set was engineered during a series of experiments run on Slovene data. These experiments are described in detail in (Ljubešić and Erjavec, 2016). Our final feature set consists of the following features:

- lowercased tokens at positions -3, -2, -1, 0, +1, +2, +3
- focus token suffixes of length 1..4
- focus token packed representation giving information whether the word consists of lowecase / uppercase letters, digits or other characters, and whether it occurs

[9] http://www.chokkan.org/software/crfsuite/

[10] https://pypi.python.org/pypi/marisa-trie

at the beginning of the sentence, e.g. `ull-START` (starts with upper-case followed by at least two lower case character at the start of the sentence) or `ddxd` (starts with a sequence of more than one digit, followed by a non-alphanumberic character, and a digit at the end)

- MSD hypotheses from the lexicon for tokens at positions -2, -1, 0, 1 and 2
- binary variable whether there is a MSD hypothesis for the focus token (added to discourage tagging unknown words with closed-class part-of-speech MSDs)

5.2. Evaluation

For testing the Croatian models we use the concatenation of the three available standard test sets, each 100 sentences in size. The test sets come from the SETimes newspaper, Wikipedia and the web. For Serbian we use the equivalent SETimes and Wikipedia test sets, all together 200 sentences in size. All test sets are available from the SETimes.HR corpus repository.[11]

Regarding corpus supervision, we experiment with two different training corpora: the SETimes.HR corpus (83,637 tokens in size, used until now for training Croatian and Serbian taggers), and the new hr500k corpus (496,989 tokens in size).

For training the Serbian models we use the Croatian corpora as (1) we currently do not have a representative manually annotated Serbian corpus at our disposal[12] and (2) previous experiments have shown that only a minor drop in accuracy should be expected from this setting (Agić et al., 2013a).

Additionally, we inspect the impact of adding the presented morphological lexicons to the tagging task.

For each setting we train both the HunPos tagger and our new CRF-based tagger.

We evaluate each system via token-level accuracy on the full morphosyntactic description (MSD, 562 labels in set.hr and 773 labels in hr500k) and the part-of-speech (POS, 12 labels in set.hr and 13 labels in hr500k).

The results for Croatian are given in Table 7. Concerning the relationship between HunPos and our CRF tagger, on full MSDs the CRF tagger consistently outperforms HunPos. While the difference between the two taggers before including the lexicon is rather small (0.44% when training on set.hr, 1.62% when training on hr500k), it becomes more significant after the inclusion of the lexicon (2.06% on set.hr, 3.23% on hr500k). This difference can be explained by the fact that the CRF-based tagger uses the lexicon during training while HunPos uses the lexicon just during annotation.

An interesting observation is that on the part-of-speech level, before including the lexicon, HunPos outperforms the CRF-based tagger. Nevertheless, in the best performing setting for both taggers (using both the lexicon and the hr500k corpus) the CRF tagger outperforms HunPos in that category as well.

tagger	lexicon	corpus	MSD	POS
HunPos	-	set.hr	84.92%	96.48%
HunPos	hrLex	set.hr	87.71%	97.88%
CRF	-	set.hr	85.36%	94.91%
CRF	hrLex	set.hr	89.77%	97.37%
HunPos	-	hr500k	89.01%	97.75%
HunPos	hrLex	hr500k	89.30%	97.86%
CRF	-	hr500k	90.63%	97.07%
CRF	hrLex	hr500k	**92.53%**	**98.11%**

Table 7: Results for Croatian morphosyntactic tagging

tagger	lexicon	corpus	MSD	POS
HunPos	-	set.hr	84.30%	96.06%
HunPos	srLex	set.hr	87.96%	97.41%
CRF	-	set.hr	84.83%	94.30%
CRF	srLex	set.hr	90.48%	97.53%
HunPos	-	hr500k	85.82%	95.94%
HunPos	srLex	hr500k	87.20%	96.65%
CRF	-	hr500k	88.34%	95.94%
CRF	srLex	500k	**92.33%**	**97.86%**

Table 8: Results for Serbian morphosyntactic tagging

Regarding the impact of the larger corpus, before including the lexicon the absolute accuracy gain is 5.27% (an error reduction of 36%) while when using the lexicon the accuracy gain, as one would expect, decreases to 2.76% (27% error reduction). The absolute accuracy gain obtained when adding the lexicon is 4.41% (30% error reduction), so, from the starting point of having only the set.hr corpus for training the tagger, by extending the training corpus to hr500k, we can observe an 0.86% better result than when including the hrLex lexicon. This follows our conclusions in (Ljubešić and Erjavec, 2016) where we show that corpus supervision is more crucial than lexicon supervision. One has to take into account that manually checking the ∼410 thousand tokens takes ∼285 linguist hours while producing a lexicon of ∼100 thousand lemmas takes around 2,000 linguist hours, so 7 times more.

Comparable experiments performed on the Serbian lexicon, Croatian corpora and the Serbian test set are given in Table 8.

The most interesting observation in this batch of experiments is that by adding the Serbian lexicon we gain a larger improvement than on Croatian data (5.65% vs. 4.41% on set.hr, 3.99% vs. 1.90% on hr500k), showing that the accuracy loss obtained by using non-native training corpora (2.29% when using hr500k and no lexicon) can almost be eliminated by adding a native lexicon (0.2% on hr500k and using a lexicon). We should stress here that throughout the experiments the Serbian test set has shown to be simpler than the Croatian test set, which can also be observed in slightly better results obtained for Serbian when using set.hr and the lexicon than on Croatian with the same settings. This is why the absolute difference in the best tagger performances for each language of 0.2% only has to be taken with caution. However, the observation that there is

[11] `https://github.com/ffnlp/sethr`

[12] We consider our Croatian corpora to be more useful for training Serbian taggers than the MulTextEast "1984" corpus of Serbian because of its specific domain.

significantly more improvement when adding the lexicon if the corpus is not native still holds.

Regarding the relationship between HunPos and the CRF-based tagger on Serbian test data, we can observe that the significant positive impact of the lexicon on the CRF-based tagger is not as present in case of HunPos (absolute accuracy gain of 3.99% vs. 1.38% on hr500k), again because HunPos does not use the lexicon during training. When comparing the best performing systems built with HunPos and the CRF-based tagger, the difference in accuracy of 5.13% (vs. a difference of 3.23% on Croatian data) becomes even more imminent. This shows for the CRF-based lexicon to be more adaptable in cases where one combines data sources with a different background, either in form of another closely related language or another register.

6. Conclusion

In this paper we have introduced two new large inflectional lexicons of Croatian and Serbian published under a very permissive license, and have shown that the lexicon content is of acceptable quality (1.9% error rate). We have also introduced a 500 thousand token training dataset for Croatian annotated on the level of morphosyntax and lemma.

We have applied the three resources to the task of morphosyntactic tagging, presenting best results on this task for both languages. While previous systems were achieving ∼85% accuracy on test sets that were frequently in-domain, we report ∼92.5% accuracy on mixed-domain test sets. We have released both the code of the tagger code and pretrained models for Croatian, Serbian and Slovene on `https://github.com/uzh/reldi/tree/master/tools/tagger`.

We have shown that our CRF-based tagger improves more when increasing corpus supervision than adding lexicon supervision. Corpora are, additionally, cheaper resources to produce. However, adding the large lexicons does push the results significantly further, a phenomenon much less observable with the HunPos tagger.

On Serbian, as we perform corpus supervision with Croatian data, the impact of adding a lexicon is even bigger than in case of Croatian, showing the increase in robustness of the tagger when adding a lexicon. The two languages are very similar and the problem of porting technologies between the two languages can often be placed under the umbrella of domain adaptation problems.

Future work on the lexicons includes a manual check of all lexicon entries, paradigm by paradigm, which should significantly lower the small noise level currently present in the lexicon.

Regarding further developments on the training corpus, a part is already annotated with Universal Dependencies and we are in the process of annotating further parts of the corpus on that level. We are working on producing Serbian training corpora as well.

7. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran) and the Swiss National Science Foundation grant IZ74Z0_160501 (ReLDI)

8. Bibliographical References

Agić, Ž. and Ljubešić, N. (2014). The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Agić, v. and Ljubešić, N. (2015). Universal Dependencies for Croatian (that work for Serbian, too). In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 1–8.

Agić, v., Ljubešić, N., and Merkler, D. (2013a). Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the Fourth Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Agić, Ž., Merkler, D., and Berović, D. (2013b). Parsing Croatian and Serbian by using Croatian dependency treebanks. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Klubička, F. and Ljubešić, N. (2014). Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of croatian. In Tomaž Erjavec et al., editors, *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*, Ljubljana, Slovenia.

Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade, Belgrade.

Ljubešić, N. and Klubička, F. (2014). {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Ljubešić, N., Stupar, M., and Jurić, T. (2012). Building named entity recognition models for croatian and slovene. In Tomaž Erjavec et al., editors, *Proceedings of the Eighth LANGUAGE TECHNOLOGIES Conference*, Ljubljana, Slovenia.

Ljubešić, N., Esplà-Gomis, M., Klubička, F., and Preradović, N. M. (2015). Predicting inflectional paradigms and lemmata of unknown words for semi-automatic expansion of morphological lexicons. In *Proceedings of Recent Advances in Natural Language Processing*.

Ljubešić, N. and Erjavec, T. (2016). Corpus vs. lexicon supervision in morphosyntactic tagging: The case of slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). `http://www.chokkan.org/software/crfsuite/`.

Šnajder, J., Bašić, B. D., and Tadić, M. (2008). Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management*, 44(5):1720–1731.

Straková, J., Straka, M., and Hajic, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 13–18.

Tadić, M. and Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Language*. ACL.

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

Vitas, D., Rehm, G., and Uszkoreit, H. (2012). *The serbian language in the digital age*. Springer.