Improving UD processing via satellite resources for morphology

Kaja Dobrovoljc
Jožef Stefan Institute
Ljubljana, Slovenia
kaja.dobrovoljc@ijs.si

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Nikola Ljubešić
Jožef Stefan Institute
Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Abstract

This paper presents the conversion of the reference language resources for Croatian and Slovenian morphology processing to UD morphological specifications. We show that the newly available training corpora and inflectional dictionaries improve the baseline `stanfordnlp` performance obtained on officially released UD datasets for lemmatization, morphology prediction and dependency parsing, illustrating the potential value of such satellite UD resources for languages with rich morphology.

1 Introduction

Many treebanks and tools are nowadays available for natural language processing tasks based on the Universal Dependencies (UD) framework, aimed at cross-linguistically consistent treebank annotation to facilitate multilingual parser development, cross-lingual learning, and parsing research (Nivre et al., 2016). As shown by the two successive CoNLL shared tasks on multilingual parsing from raw text to UD (Zeman et al., 2017, 2018), existing UD systems achieve state-of-the-art results both in terms of dependency parsing and lower levels of grammatical annotation.

However, in addition to the officially released UD treebanks with complete syntactic and morphological annotations, the rapidly emerging UD tools would benefit from other language resources, as well. This is especially true for morphological annotation (lemmatization, PoS tagging and morphological feature prediction), as many languages employ much larger morphology-annotated corpora than the costly (sub)corpora annotated for syntax, as well as morphological lexicons, essential for high-quality processing of languages with complex morphology.

Examples of such cases are Croatian and Slovenian, two South Slavic languages with rich inflection. Their official UD releases include the conversions of the largest syntactically annotated corpora available for each language (Agić and Ljubešić, 2015; Dobrovoljc et al., 2017a), however, other manually created resources, such as the larger morphologically annotated corpora (Ljubešić et al., 2018b; Krek et al., 2019) and inflectional lexicons (Ljubešić, 2019; Dobrovoljc et al., 2019), have also been developed to support the development of related NLP tools (Ljubešić and Erjavec, 2016; Grčar et al., 2012) in the past.

The aim of this paper is to present the conversion of these resources to the UD formalism and explore their potential contribution to the state-of-the-art in UD processing for both languages, from lemmatization to morphology and syntax prediction. Using the `stanfordnlp` tool, we investigate the impact of newly available data on all three tasks by (1) retraining the tagging and lemmatization models on larger training sets and (2) performing a simple lexicon lookup intervention in the lemmatization procedure.

This paper is structured as follows. We first briefly describe the creation and the content of the newly released resources for both languages in Section 2, followed by the presentation of the experiments for their evaluation in Section 3. We present the corresponding results in Section 4 and conclude in Section 5 by a short discussion of their wider implications for related UD languages and the UD community in general.

2 Extending the resources for UD morphology

This section describes the development, the content and the availability of the extended UD resources for Slovenian and Croatian, namely the larger training sets for UD morphology (the ssj500k and hr500k cor-

pora) and the large-scale UD-compliant lexicons of inflected forms (Sloleks and hrLex). Given the methodological differences in resource development for both languages due to divergent project frameworks and scopes, we present the resources by language rather than type. However, a brief quantitative overview and comparison is given at the end of the section.

2.1 Slovenian resources

Both the ssj500k training corpus (Erjavec et al., 2010) and the Sloleks lexicon of inflected forms (Dobrovoljc et al., 2017b) adopt the JOS morphosyntactic annotation scheme (Erjavec and Krek, 2008), compatible with MULTEXT-East morphosyntactic specifications (Erjavec, 2012), which define the part-of-speech categories for Slovene, their morphological features (attributes) and values, and their mapping to morphosyntactic descriptions (MSDs).[1] An automatic rule-based mapping from JOS to UD part-of-speech tags and features had already been developed as part of the original Slovenian UD Treebank conversion from the syntactically annotated subset of the ssj500k corpus (Dobrovoljc et al., 2017a), with the conversion scripts now publicly available at the CLARIN.SI GitHub repository.[2]

The large majority of conversion rules for morphology are direct mappings of specific categories – e.g. conversion of JOS numerals (`M`) with `Form=letter` and `Type=ordinal` to UD adjectives (`ADJ`) with feature `NumType=Ord` – making them directly applicable for converting any language resource with JOS morphosyntactic annotations, such as the resources presented in this paper. The only exception are the two rules involving predefined lists of JOS pronouns and adverbs to be converted to UD determiners (e.g. *ta* 'this' or *veliko* 'many'), which have been updated so as to cover the previously unknown vocabulary emerging from ssj500k and Sloleks (i.e. adding 135 new lemmas to the list of UD determiners).

2.1.1 ssj500k corpus

The ssj500k training corpus is the largest training corpus for Slovenian, with approx. 500,000 tokens manually annotated on the levels of tokenization, segmentation, lemmatization and morphosyntactic tagging. Variously-sized ssj500k subsets have also been annotated for other linguistic layers, namely named entities, JOS dependency syntax, semantic roles, verbal multi-word expressions and Universal Dependencies.

To extend UD morphology annotations to the entire ssj500k corpus, v2.1 of the corpus (Krek et al., 2018) was converted using the pipeline referenced above. Specifically, the script `tei2ud.xsl` takes the original XML TEI format as input, converts it to a CONLL-like tabular format with English JOS tags, features and dependencies, followed by the conversion to the standardized CONLL-U file with UD PoS and morphological features. This second step is performed by the `jos2ud.pl` script, which takes two mapping files as parameters, one for the PoS mapping (`jos2ud-pos.tbl`), and the other for feature mapping (`jos2ud-features.tbl`).

For occurrences of the verb *biti* ('to be') – the only instance of the PoS mapping depending on syntactic role – an additional set of scripts is applied (`add-biti-*.pl`) to disambiguate between the auxiliary, copula (both `AUX` in UD) and other (`VERB` in UD) usages of this verb, which are always labelled as an auxiliary verb in JOS. In contrast to the occurrences within syntactic trees enabling rule-based disambiguation and the unambiguous occurrences of *biti* preceding verbal participles (and potentially intervening pronous, adverbs, particles or conjunctions), the remaining 11,925 *biti* tokens in ssj500k have been disambiguated manually. This was performed by trained native speakers, with two annotators per decision and a third one in case of competing annotations (93.9% agreement, Cohen's Kappa 0.78).

The resulting ssj500k corpus with UD PoS tags, morphological features and their values has been released as part of ssj500k release v2.2 (Krek et al., 2019) under CC BY-NC-SA 4.0. In addition to the CONLL-U format, in which underscores have been inserted where the dependency annotations are missing, the information on UD morphology and syntax has also been added to the original TEI XML format with other types of annotation and metainformation, as illustrated in Figure 1.

The sentence element (`<s>`) contains words (`<w>`), punctuation symbols (`<pc>`) and whitespace (`<c>`), as well as segments (`<seg>`) for annotating spans of tokens, and link groups (`<linkGrp>`) for annotating

[1] The latest (Version 6) MULTEXT-East multilingual morphosyntactic specifications are available at `http://nl.ijs.si/ME/V6/` and being developed at `https://github.com/clarinsi/mte-msd`.

[2] `https://github.com/clarinsi/jos2ud`

```
<s xml:id="ssj1.1.2">
   <w ana="mte:Ncmsn" msd="UposTag=NOUN|Case=Nom|Gender=Masc|Number=Sing"
      lemma="dogodek" xml:id="ssj1.1.2.t1">Dogodek</w><c> </c>
   <w ana="mte:Sl" msd="UposTag=ADP|Case=Loc"
      lemma="v" xml:id="ssj1.1.2.t2">v</w><c> </c>
   <seg type="name" subtype="loc">
      <w ana="mte:Npmsl" msd="UposTag=PROPN|Case=Loc|Gender=Masc|Number=Sing"
      lemma="Ankaran" xml:id="ssj1.1.2.t3">Ankaranu</w>
   </seg><c> </c>
   <w ana="mte:Va-r3s-n"
      msd="UposTag=AUX|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|
      VerbForm=Fin"
      lemma="biti" xml:id="ssj1.1.2.t4">je</w><c> </c>
   <w ana="mte:Va-p-sf" msd="UposTag=AUX|Gender=Fem|Number=Sing|VerbForm=Part"
      lemma="biti" xml:id="ssj1.1.2.t5">bila</w><c> </c>
   <w ana="mte:Agpfsn" msd="UposTag=ADJ|Case=Nom|Degree=Pos|Gender=Fem|Number=Sing"
      lemma="dramatičen" xml:id="ssj1.1.2.t6">dramatična</w><c> </c>
   <w ana="mte:Ncfsn" msd="UposTag=NOUN|Case=Nom|Gender=Fem|Number=Sing"
      lemma="nesreča" xml:id="ssj1.1.2.t7">nesreča</w>
   <pc ana="mte:Z" msd="UposTag=PUNCT" xml:id="ssj1.1.2.t8">.</pc>
   <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="UD-SYN">
     <link ana="ud-syn:root" target="#ssj1.1.2 #ssj1.1.2.t1"/>
     <link ana="ud-syn:case" target="#ssj1.1.2.t3 #ssj1.1.2.t2"/>
     <link ana="ud-syn:nmod" target="#ssj1.1.2.t1 #ssj1.1.2.t3"/>
     <link ana="ud-syn:aux" target="#ssj1.1.2.t1 #ssj1.1.2.t4"/>
     <link ana="ud-syn:cop" target="#ssj1.1.2.t1 #ssj1.1.2.t5"/>
     <link ana="ud-syn:amod" target="#ssj1.1.2.t7 #ssj1.1.2.t6"/>
     <link ana="ud-syn:nsubj" target="#ssj1.1.2.t1 #ssj1.1.2.t7"/>
     <link ana="ud-syn:punct" target="#ssj1.1.2.t1 #ssj1.1.2.t8"/>
   </linkGrp>
   <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="JOS-SYN">
      <link ana="jos-syn:Atr" target="#ssj1.1.2.t5 #ssj1.1.2.t1"/>
      <link ana="jos-syn:Atr" target="#ssj1.1.2.t3 #ssj1.1.2.t2"/>
      <link ana="jos-syn:Atr" target="#ssj1.1.2.t1 #ssj1.1.2.t3"/>
      <link ana="jos-syn:PPart" target="#ssj1.1.2.t5 #ssj1.1.2.t4"/>
      <link ana="jos-syn:Root" target="#ssj1.1.2 #ssj1.1.2.t5"/>
      <link ana="jos-syn:Atr" target="#ssj1.1.2.t7 #ssj1.1.2.t6"/>
      <link ana="jos-syn:Sb" target="#ssj1.1.2.t5 #ssj1.1.2.t7"/>
      <link ana="jos-syn:Root" target="#ssj1.1.2 #ssj1.1.2.t8"/>
   </linkGrp>
   <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="SRL">
      <link ana="srl:ACT" target="#ssj1.1.2.t5 #ssj1.1.2.t1"/>
      <link ana="srl:PAT" target="#ssj1.1.2.t5 #ssj1.1.2.t7"/>
   </linkGrp>
</s>
```

Figure 1: The full TEI encoding of the sentence *Dogodek v Ankaranu je bil dramatična nesreča.* ('The incident in Ankaran was a dramatic accident.').

links between tokens. The words contain annotation on their JOS (MULTEXT-East) morphosyntactic description (the `@ana` attribute), as well as the Universal Dependencies morphosyntactic features (`@msd`), the lemma of words (`@lemma`) and the ID of each token (`@xml:id`). The fact that a segment denotes a named entity is signaled by `@type="name"`, and the type of the named entity by the `@subtype` attribute. The Universal Dependencies syntactic relations are encoded in the `<linkGrp type="UD-SYN">` element, where the individual links give the head and argument of the relation, which is encoded in the `@ana` attribute. Note that the sentence identifier serves as a proxy for the virtual syntactic root of the sentence tree. Similarly, the JOS syntactic relations are encoded in the `<linkGrp type="JOS-SYN">` element. Finally, the semantic role relations are encoded in the `<linkGrp type="SRL">` element.

2.1.2 Sloleks morphological lexicon

The Sloleks morphological lexicon is the largest manually created collection of inflected forms in Slovenian, consisting of 2,792,003 inflected forms and 100,805 lemmas, with each inflected form bearing information on its lemma, grammatical features, pronunciation and frequency of usage. Version 1.2 of the lexicon (Dobrovoljc et al., 2015) has been converted using the same JOS-to-UD conversion script, which allows switching between corpus and lexicon mode. The converted lexicon with UD PoS tags (UPOS), features and values (FEATS) has been released as part of the Sloleks release 2.0 (Dobrovoljc et al., 2019) under CC BY-NC-SA 4.0, in the form of a tab-separated file listing the inflected form, its lemma, JOS MSD tag, frequency of usage, JOS PoS and features, and UD PoS and features. The mapping to the original Sloleks release in LMF XML encoding with several additional layers of information, such as pronunciation, is not explicit, but can be reproduced based on unique combinations of the given features.

2.2 Croatian resources

The hr500k training corpus (Ljubešić et al., 2018b) and the hrLex inflectional lexicon (Ljubešić, 2019) were developed on the margins of many projects, with the ReLDI project[3] giving the final push for their consolidation and publication.

The enrichment of these resources with UD information was format-wise very similar to that of the Slovenian resources described in Section 2.1, with (1) differences in the mapping of MULTEXT-East morphosyntactic annotations to the Universal Part-of-Speech (UPOS) and morphological features (FEATS) due to a slightly different tagset for Croatian and (2) no mappings performed on the dependency syntax level, as the corpus was manually annotated with the UD dependency syntax layer.

2.2.1 hr500k training corpus

The hr500k training corpus contains tokens manually annotated on the levels of tokenisation, sentence segmentation, morphosyntactic tagging, lemmatization and named entities. About half of the corpus is also manually annotated with UD syntactic dependencies. Furthermore, about a fifth of the corpus is annotated with semantic role labels. This corpus is considered to be the reference training corpus for Croatian. The details on the content of the corpus are described in Ljubešić et al. (2018a).

The morphosyntactic layer of the corpus was initially annotated with the MULTEXT-East morphosyntactic specifications (Erjavec, 2012) and the mapping to the UPOS and FEATS layers was performed semi-automatically, with the automatic part consisting of (1) applying an explicit mapping between MULTEXT-East tags and UPOS and FEATS tags[4] and (2) fallback to additional rules for pronouns and determiners, adverbs, numbers and the negated auxiliary.[5] The only non-automatic part of the mapping was the resolution of the category of abbreviations from MULTEXT-East to the corresponding parts-of-speech.

The resulting hr500k corpus was part of the initial release of hr500k (v1.0) and was published under CC BY-SA 4.0 (Ljubešić et al., 2018b).

2.2.2 hrLex inflectional lexicon

The hrLex inflectional lexicon (Ljubešić, 2019) is currently the largest inflectional lexicon of Croatian. The process of semi-automatically building the hrLex inflectional lexicon is described in Ljubešić et al. (2016).

[3] `https://reldi.spur.uzh.ch`

[4] `https://github.com/nljubesi/hr500k/blob/master/mte5-udv2.mapping`

[5] `https://github.com/vukbatanovic/SETimes.SR/blob/master/msd_mapper.py`

The mapping of the MULTEXT-East tags that were initially present in the lexicon to the UPOS and FEATS layers was performed by applying the mapping that was used to map the hr500k training corpus to these layers, without the need for the manual mapping.

The UD information became part of the hrLex lexicon with version 1.3 (Ljubešić, 2019), when the lexicon was published under the CC BY-SA 4.0 license. The lexicon is published as a tab-separated file listing the inflected form, its lemma, MULTEXT-East tag, MULTEXT-East morphological features, UPOS, FEATS, and the absolute and relative (per-million) frequency of usage in the hrWaC corpus (Ljubešić and Klubička, 2016).

2.3 Quantitative overview

This section gives a quantitative overview of the newly available resources for both languages, to illustrate their morphological complexity and the importance of the corresponding disambiguation in the process of morphological annotation and lemmatization (Section 3).

Table 1 shows, for the Slovene and Croatian corpora, first the number of tokens and types, where the latter is taken to be triplets consisting of the lower-cased wordform (i.e. token), lemma, and the MULTEXT-East XPOS (giving both PoS and features). This is followed by the numbers of each of the individual members of the triplet. As can be seen, both corpora have approximately half a million tokens, and somewhat under 100,000 lexical types, with the Croatian resource being somewhat smaller and having a poorer lexicon, most likely because of its more restricted variety of source texts. The Croatian corpus also uses almost half less tags, however, this follows from the overall smaller number of defined tags, as will be shown in the discussion of the lexicon.

Next are shown the numbers of out-of-vocabulary tokens and types against the two lexicons, Sloleks and hrLex, but not taking into account punctuation, which is not part of the lexicon. The Croatian corpus has almost twice as many OOV types and tokens, which is due to the construction of the Slovene lexicon, discussed below.

The last column gives the type ambiguity in the corpora, i.e. on the average, how many different interpretations (lemmas or tags) does each distinct wordform have. In both cases the number is very similar, $5/4$. This means that, on average, each fourth word will have two interpretations, which is a simplified view of ambiguity, as some distinct wordforms have more than two interpretations.

	Tokens	**Types**	**Wforms**	**Lemmas**	**Tags**	**OOV types**	**OOV toks**	**Ambig.**
ssj500k	586,248	98,641	78,707	38,818	1,304	5.26%	18.33%	1.25
hr500k	506,457	84,789	66,797	34,321	768	9.70%	27.17%	1.27

Table 1: Size of newly available corpora for UD morphology.

Table 2 gives a quantitative overview of the two lexicons. The number of entries is the number of wordform / lemma / tag triplets, and the next three columns give, as with the corpora, the individual numbers of wordforms, lemmas and morphosyntactic tags. As can be seen, hrLex is almost twice as large as Sloleks, however, it does distinguish only about half the tags compared to Slovene. This is mostly due to the tags related to the dual number in Slovene and a very fine-grained typology of Slovene pronouns, which account for almost half of the tagset. This is also evidenced by the number of tags used on the Slovene corpus (1,304), which, although much larger than for Croatian, is much smaller than the lexicon inventory.

The last column gives the ambiguity in the lexicon, i.e. how many different interpretations in terms of lemma and tag does, on the average, one wordform have. As can be seen, this number is over three in both cases, with the Croatian ambiguity being significantly higher; we discuss the reasons below. It can also be noticed that the lexicon ambiguity is in both cases much greater than in the corpora, which is due to the fact that the lexicons contain the complete inflectional paradigms, although some of its word forms are rarely present in the corpora.

Figure 2 gives — on a logarithmic scale — the lemma sizes of the two lexicons by the UD part-of-speech. The most striking features are the significantly greater number of adverbs, adjectives and proper nouns of the Croatian lexicon. This stems from the automatic inclusion of adverbs derived from adjectives,

	Entries	Wforms	Lemmas	Tags	Ambig.
Sloleks	2,792,003	921,869	96,593	1,900	3.03
hrLex	6,427,709	1,697,943	164,206	900	3.79

Table 2: Size of newly available lexica for UD morphology.

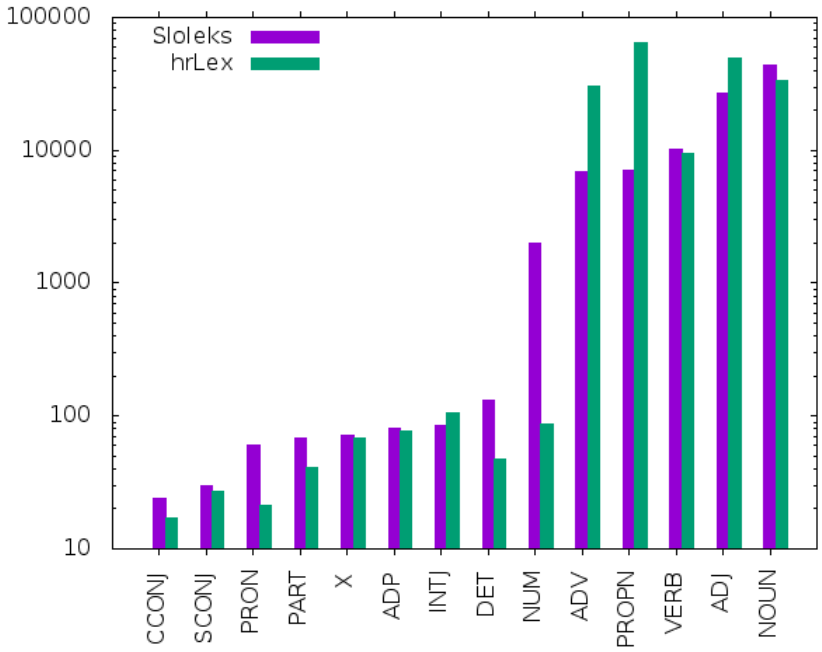

Figure 2: Size of lexicons by UD part-of-speech

and possessive adjectives derived from nouns in hrLex, and, of course, the preference given to including a large number of proper nouns. In contrast, Sloleks was constructed purely on the basis of quantitative criteria, i.e. it includes 100,000 lemmas which had the highest frequency in the large 600-million-word corpus FidaPLUS, and the majority of tokens occurring in the ssj500k corpus, which also explains its lower OOV rates in Figure 1. In any case, the different principles in the creation of the lexicons account for the larger size of hrLex lexicon and also explain the large difference in the ambiguity of the two lexicons, as possessive adjectives and proper nouns have a somewhat higher ambiguity than the remainder of the lexicon: possessive adjectives have an ambiguity of 4.68, and proper nouns of 3.78.

3 Experiment setup

3.1 Tool

We perform experiments on morphosyntactic tagging, lemmatization and dependency parsing via the `stanfordnlp` tool, one of the best-performing systems in the CoNLL shared task in 2018 (Zeman et al., 2018) with code released and a vivid development community.[6] The details on the implementation of the tool are given in (Qi et al., 2018). The tool assumes that morphosyntactic tagging is performed first, producing the UPOS and FEATS annotation layer. Next, lemmatization is performed by using the UPOS (but not FEATS) predictions. Finally, parsing is performed by exploiting all previously predicted layers (UPOS, FEATS and LEMMA). We investigate the impact of additional data on all three tasks by (1) retraining morphosyntactic tagging and lemmatization models with more data and (2) performing a simple intervention in the lemmatization procedure so that the lexicon lookup is not performed over the training data only, but the external inflectional lexicon as well.

[6] `https://github.com/stanfordnlp/stanfordnlp`

3.2 Data split

The `babushka-bench`[7] is a benchmarking platform currently used for three South Slavic languages, namely Slovenian, Croatian and Serbian (Ljubešić and Dobrovoljc, 2019). The name of the benchmarking platform comes from the idea that similar splits of data may be performed for various levels of linguistic annotation in a dataset, regardless of the fact that not all data is annotated on all linguistic levels. Each dataset is split (with a fixed random seed) via a pseudorandom function so that 80% of the data is allocated for train, while 10% is allocated to dev and 10% to test. If the dataset is split on a linguistic level which is not covered in the whole dataset, instances that do not have that level of annotation are simply discarded. What such a split enables, which becomes evident in this research already, is that it is safe to use training data from the split on the morphosyntactic level (which is applied on the whole dataset) and use the resulting model on experiments on the dependency syntax level (which is available in less than half of the dataset) without fears of data spillage between train, dev and test (e.g. the test set for parsing containing sentences that are used for training the tagger, therefore applying the tagger on the parsing test data would produce unrealistically good tags, thereby unrealistically improve parsing). The size of the data splits, both on the morphosyntactic (i.e. UD morphology annotations) and the dependency syntax (i.e. full UD annotations) levels, used in this research is given in Table 3.[8]

	ssj500k	**sl UD**	**hr500k**	**hr UD**
train	474,322	110,711	415,328	165,989
dev	62,967	16,589	39,765	14,184
test	48,959	13,370	51,364	16,855
Σ	586,248	140,670	506,457	197,028

Table 3: The benchmarking data split of the ssj500k and hr500k corpora and their officially released UD subsets.

3.3 Training and evaluation

The experiments in this paper are organised in two parts: the experiments with an extended training corpus on the level of morphosyntax and lemma and the experiments on adding an inflectional lexicon to the lemmatization process.[9]

While we perform experiments on the levels of morphosyntax, lemma and dependency syntax, we use gold segmentation to simplify our experiments as different tokenisers and sentence splitters are available for the two languages in question. Performing different preprocessing on the two languages would blur our experiments. On the other hand, applying the out-of-the box segmentation of `stanfordnlp` would produce results that are detrimental to those of our rule-based tokenizers and sentence splitters.[10] Overall, our previous experiments show that true segmentation deteriorates the results slightly on all levels of annotation, but that relations between results of different systems or setups hold regardless of whether gold or true segmentation is used.

When performing training and evaluation on levels of lemmatization and dependency syntax, we preannotate all the three data portions (train, dev and test) with the models from the upstream levels. We therefore apply morphosyntactic models on the data to be used for training and evaluating lemmatization, and we apply morphosyntactic tagging and lemmatization before training and evaluating dependency parsing models. While it is to be expected that training and applying the models on the training data will give an

[7] `https://github.com/clarinsi/babushka-bench`

[8] For both languages, the babushka split of data with full UD annotations differs from the official UD data releases, which are advised not to change across UD releases. However, baseline experiment results for both data split versions remain comparable (see Section 4).

[9] We do not consider improving morphosyntactic annotation via the inflectional lexicon in this paper as initial experiments have shown that various approaches to simple application of the inflectional lexicon (via lookup) do not yield any improvements. Exploiting the inflectional lexicon, probably while training the morphosyntactic annotation model, is left for future work.

[10] Readers interested in the comparison between the various segmenters should investigate the results published at `https://github.com/clarinsi/babushka-bench`

unrealistically good automatic annotation of the training data, our intuition is that, given that development data can be considered realistically annotated, the final impact of this simplifying solution (jack-knifing, i.e. annotating the training data via cross-validation would be an alternative) on the quality of annotation of the test (or any other) data will be minimal, if any. Simply preannotating training data with the model trained on that same data was also the approach taken by the developers of `stanfordnlp` during the CoNLL 2018 shared task (Qi et al., 2018).

The experiments on using a larger dataset for training the morphosyntactic tagging and lemmatization models, for which we expect to have a positive impact on the parsing quality, are split into two main parts: (1) training and evaluating morphosyntactic tagging and lemmatization models on the UD data and on all the available data, and (2) applying both models as pre-processing for training and evaluating models for dependency parsing.

The experiments on using the inflectional lexicon for improving lemmatization by extending the lookup method on the external lexicon, consist, similarly, of the experiments on training and evaluating the lemmatization models based on UD and all the available data, both with and without the lexicon, and inspecting the impact of the improved lemmatization on the parsing quality.

We evaluate all approaches with the evaluation script of the CoNLL 2018 shared task (Zeman et al., 2018), reporting F1 on all relevant levels, these being LEMMA, UPOS, XPOS, FEATS scores for morphology. For dependency syntax, the standard unlabelled (UAS) and labelled (LAS) attachment scores are complemented with the recently proposed morphology-aware labelled attachment score (MLAS), which also takes part-of-speech tags and morphological features into account and treats function words as features of content words, and bi-lexical dependency score (BLEX), which is similar to MLAS, but also incorporates lemmatization. For evaluation, we use only the UD portions of the test datasets to keep the numbers obtained on the UD data and the extended data as comparable as possible.

4 Results

The results, summarized in Tables 4 and 5, show the improvements in baseline `stanfordnlp` lemmatization, tagging and parsing performance for Croatian and Slovenian, based on the integration of the newly available training datasets for UD tagging and lemmatization (Section 4.1) and large-scale inflectional dictionaries (Section 4.2) for lemmatization.

4.1 Training corpus

Table 4 shows that re-training the lemmatization and tagging models on larger UD training sets (the ssj500k and hr500k corpora) improves the baseline performance obtained on officially released UD data[11] for both languages and for all evaluation metrics selected. In particular, the largest improvements are observed for lemmatization (+1.56pp for Slovenian and +0.91 for Croatian), language-specific JOS MSD tagging (XPOS) (+1.35 / +0.52) and universal morphological feature prediction (+1.28 / +0.53). The impact of a threefold training set increase is much less pronounced for universal PoS categories on an absolute scale (+0.24 / +0.14), but also on a relative one (15.6% vs. 31% relative error reduction on Slovenian UPOS vs. XPOS), which shows greater benefits of additional training data for the more complex layers of detailed morphosyntactic description.

As expected, retraining the parsing models on data with improved (predicted) morphology, benefits the parsing performance, as well, esp. for the morphology-sensitive scores MLAS (+1.98 / +1.34) and BLEX (+2.92 / +2.11). For the standard LAS score, the improvements amount to approx. 0.7pp for both languages. For all selected metrics, the improvements for Slovenian data are higher in comparison to Croatian, which is understandable given the differences in training data increase, i.e. a 4.3-fold increase for Slovenian and a 2.5-fold increase for Croatian (Figure 3).

[11] The baseline `stanfordnlp` performance on `babushka-bench` split is similar to that on official splits, as reported in `https://stanfordnlp.github.io/stanfordnlp/performance.html`, with the exception of FEATS prediction for Croatian, where official UD data has a specifically hard test set in comparison to the training data.

	sl UD	sl 500	hr UD	hr 500
LEMMA	95.88	97.44	95.30	96.21
UPOS	98.45	98.69	97.91	98.05
XPOS	95.65	97.00	94.60	95.12
FEATS	95.95	97.23	95.13	95.66
UAS	93.40	93.72	90.22	90.76
LAS	91.62	92.28	85.30	86.00
MLAS	84.24	86.22	75.54	76.88
BLEX	84.04	86.96	76.45	78.56

Table 4: Improvements in baseline `stanfordnlp` lemmatization, tagging and parsing performance for Croatian and Slovenian through a larger training set for UD morphology.

4.2 Lexicon of inflected forms

The results in Table 5 show that introducing large-scale morphological dictionaries (the Sloleks and hrLex lexicons) through a simple dictionary lookup significantly improves the performance of the lemmatization models trained on official UD training data alone (+2.6pp for Slovenian / +1.94 for Croatian). Noticeable improvements are also observed in comparison to the lemmatization models trained on the two larger training sets (+1.45 / +1.08), illustrating the overall benefits of morphological dictionaries in lemmatizing morphologically rich languages. As expected, the improvements are higher for Slovenian, given the larger number of OOV tokens in hr500k in comparison to ssj500k (Table 1).

Nevertheless, with the exception of lemmatization-aware BLEX score (+2.05 / +1.45), the gains in lemmatization just mildly transfer to parsing scores, giving small or no improvements. Interestingly, the improvements of parsing by improving lemmatization are consistent and overall more visible when morphosyntax and lemmatization are not improved with the extended training corpus (columns sl UD (+lex) and hr UD (+lex)), showing a saturation effect when both new datasets are exploited.

	sl UD	+ lex	sl 500	+ lex	hr UD	+lex	hr 500	+ lex
LEMMA	95.88	98.48	97.44	98.89	95.30	97.24	96.21	97.29
UAS	93.40	93.43	93.72	93.72	90.22	90.53	90.76	90.44
LAS	91.62	91.75	92.28	92.27	85.30	85.81	86.00	85.85
MLAS	84.24	84.34	86.22	86.05	75.54	76.16	76.88	76.83
BLEX	84.04	88.00	86.96	89.01	76.45	79.60	78.56	80.04

Table 5: Improvements in baseline `stanfordnlp` lemmatization, tagging and parsing performance for Croatian and Slovenian through a simple lexicon lookup for lemmatization.

5 Conclusion

This paper presented the development and the content of newly available UD-compliant training corpora and inflectional dictionaries for Slovenian and Croatian morphology processing, and illustrated their potential value to state-of-the-art tools for UD processing. Specifically, our results show that both types of resources substantially improve the baseline lemmatization, PoS tagging and morphological feature prediction performance obtained on officially released UD datasets for each language, contributing to slight improvements in dependency parsing performance, as well.

These results give important insight into the possible improvements of future text-processing tools for Slovenian, Croatian and other morphologically rich languages, where large-scale manually annotated corpora and morphological dictionaries remain relevant resources in neural-based architectures, as well. At the same time, they also raise a general question for the wider community on the optimal format, distribution and documentation of such satellite UD resources, which are likely to exist in many different languages or

can eventually emerge from other similar initiatives aimed at cross-lingual annotation of specific linguistic layers (Kirov et al., 2016; Petrov et al., 2012).

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency through the research core funding no. P6-0411 (*Language resources and technologies for Slovene language*), the research project no. J6-8256 (*New grammar of contemporary standard Slovene: sources and methods*) and the Slovenian research infrastructure CLARIN.SI.

References

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 1–8, Hissar, Bulgaria. INCOMA Ltd. Shoumen, Bulgaria.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017a. The Universal Dependencies Treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*, pages 33–38.

Kaja Dobrovoljc, Simon Krek, and Tomaž Erjavec. 2017b. The Sloleks Morphological Lexicon and its Future Development. In Vojko Gorjanc, Polona Gantar, Izok Kosem, and Simon Krek, editors, *Dictionary of Modern Slovene: Problems and Solutions*, pages 42–63. Ljubljana University Press: Faculty of Arts.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, and Miro Romih. 2015. *Morphological lexicon Sloleks 1.2*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1039`.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik, and Marko Robnik-Šikonja. 2019. *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1230`.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

Tomaž Erjavec and Simon Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. In *LREC 2008*.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. (Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene). In *Proceedings of the 8th Language Technologies Conference*, volume C, pages 89–94, Ljubljana, Slovenia. IJS.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2019. *Training corpus ssj500k 2.2*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1210`.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2018. *Training corpus ssj500k 2.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1181.

Nikola Ljubešić and Filip Klubička. 2016. *Croatian web corpus hrWaC 2.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1064.

Nikola Ljubešić. 2019. *Inflectional lexicon hrLex 1.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1232.

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018a. hr500k–a reference training corpus of croatian. In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT - DH 2018)*, pages 154–161, Ljubljana, Slovenia.

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018b. *Training corpus hr500k 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1183.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, BSNLP@ACL 2019*.

Nikola Ljubešić and Tomaž Erjavec. 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 2016)*, Paris, France. European Language Resources Association (ELRA).

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman et al. 2017. CoNLL 2017 Shared Task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.