Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 15–22
Marseille, 24 June 2022

Extending the SSJ Universal Dependencies Treebank for Slovenian: Was it Worth it?

Kaja Dobrovoljc, Nikola Ljubešić
University of Ljubljana
Jozef Stefan Institute
kaja.dobrovoljc@ff.uni-lj.si
nikola.ljubesic@ijs.si

Abstract

This paper presents the creation and the evaluation of a new version of the reference SSJ Universal Dependencies Treebank for Slovenian, which has been substantially improved and extended to almost double the original size. The process was based on the initial revision and documentation of the language-specific UD annotation guidelines for Slovenian and the corresponding modification of the original SSJ annotations, followed by a two-stage annotation campaign, in which two new subsets have been added, the previously unreleased sentences from the ssj500k corpus and the Slovenian subset of the ELEXIS parallel corpus. The annotation campaign resulted in an extended version of the SSJ UD treebank with 5,435 newly added sentences comprising of 126,427 tokens. To evaluate the potential benefits of this data increase for Slovenian dependency parsing, we compared the performance of the classla-stanza dependency parser trained on the old and the new SSJ data when evaluated on the new SSJ test set and its subsets. Our results show an increase of LAS performance in general, especially for previously under-represented syntactic phenomena, such as lists, elliptical constructions and appositions, but also confirm the distinct nature of the two newly added subsets and the diversification of the SSJ treebank as a whole.

Keywords: Slovenian, treebanks, dependency syntax, dependency parsing, Universal Dependencies, annotation guidelines, annotation campaign, data evaluation

1. Introduction

Manually annotated language data are essential to the development and evaluation of natural language processing tools. For syntactic analysis in particular, these mostly involve parsed corpora (treebanks), in which surface word forms bear additional information on their morphological and syntactic characteristics with the structure of a sentence described as a tree-like graph.

To overcome the various drawbacks rising from the multitude and heterogeneity of treebank annotation schemes, especially in the field of multilingual parser development, cross-lingual learning and research on language typology, the Universal Dependencies (UD) initiative (De Marneffe et al., 2021; Nivre et al., 2016) proposed a universal inventory of grammatical categories and guidelines for their application to facilitate consistent annotation of similar constructions across languages.

As of the latest release (Zeman and others, 2022), the UD scheme has been applied to more than 200 treebanks in over 130 languages and has contributed to important scientific advances in natural language processing and linguistics alike. This includes the reference SSJ treebank for written Slovenian (Dobrovoljc et al., 2017), which has been used in modelling several state-of-the-art parsing tools worldwide (Zeman et al., 2018). The treebank, first released in UD v1.2 in 2015, included 8,000 parsed sentences comprising of 140,670 words, placing it in the top third of UD treebanks ranked according to data size.

Within the project Development of Slovene in a Digital Environment (DSDE)[1] aimed at meeting the needs for computational tools and services in the field of language technologies for Slovenian, more than 5,000 new sentences have been added to the SSJ treebank to increase the size of manually annotated training data and thus encourage further advances in the field of Slovenian language technology.

In this paper, we present the results of this latest activity by describing the creation of the new version of the SSJ Universal Dependencies Treebank for Slovenian, which has been substantially improved both in terms of size and the quality of annotations. After a brief presentation of the original version of the treebank in Section 2, we present the extensively documented and slightly revised language-specific UD guidelines for Slovenian in (Section 3) which were implemented to the original treebank (Section 4) and used in the subsequent two-stage annotation campaign described in Section 5. We then evaluate the NLP relevance of the resulting dataset by comparing the performance of a dependency parsing tool trained on both versions of the SSJ treebank in Section 6, and conclude with a short discussion on whether our results justify the labour-intensive data extension typical of treebank annotation in general (Section 7).

2. Original SSJ UD Treebank

The original SSJ UD treebank has been created by a semi-automatic conversion from the reference ssj500k training corpus for Slovenian (Krek et al., 2020b),

[1] https://slovenscina.eu/en

a balanced collection of texts sampled from the FidaPLUS corpus (Arhar Holdt, 2007), a predecessor of the 1-billion-word Gigafida reference corpus of contemporary written Slovene (Krek et al., 2020a). The ssj500k corpus includes fiction, non-fiction and periodical texts dating from 1990 to 2000, which have been manually annotated on various levels of linguistic annotation (Krek et al., 2020b), including lemmatization, morphosyntactic tagging and dependency parsing in accordance with the JOS annotation scheme (Erjavec et al., 2010).

The ssj500k conversion from JOS to UD was based on a broad set of mapping rules for all three annotation layers (part-of-speech categories, morphological features and dependency relations),[2] the conversion to UD dependencies required highly fine-grained rules given the several significant distinctions between both annotation schemes (Dobrovoljc et al., 2017), including a much more detailed set of dependency relations in UD (37 labels) in comparison to JOS (10 labels).

As a result, the full ssj500k corpus was automatically converted to UD part-of-speech categories and morphological features with only the instances of the verb *biti* 'be' requiring manual disambiguation (Dobrovoljc et al., 2019) between *AUX* and *VERB* part-of-speech tags. On the other hand, due to the limited coverage of the the mapping rules for syntax, not all JOS-parsed sentences could be converted automatically, especially those exhibiting complex or rare phenomena pertaining to clausal coordination, juxtaposition and predicate ellipsis.

Consequently, only around two thirds of the 13,411 JOS-parsed sentences in ssj500k have been fully converted to UD, which resulted in the original SSJ UD treebank containing 8,000 sentences and 140,670 tokens. Despite the continuous improvements of the SSJ UD annotations since its first release in 2015, the size of the dataset remained unchanged. The 3,411 unreleased partially converted sentences from ssj500k were thus the obvious starting point for the recent extension of the SSJ UD dataset for Slovenian, as described in Section 5.1.

3. Slovenian UD Guidelines Revision and Documentation

With the exception of the online language-specific guidelines for UD morphology annotation published with the initial SSJ release (pertaining to the now obsolete Version 1 of the UD guidelines (Nivre et al., 2016)), the guidelines for Slovenian UD dependency annotation have only been documented implicitly – in the form of the rule-based conversion scripts from JOS to UD annotations (Section 2) and the resulting SSJ dataset. To bridge this gap and provide the necessary documentation in support of both annotation and exploration of Slovenian UD data, the official Slovenian UD guidelines have now been exhaustively documented for all layers of annotation, by describing the general annotation guidelines and its application to specific constructions in Slovenian.

In the process, a few changes to the original UD annotation principles for Slovenian were also introduced to make them better compliant with the universal guidelines and the annotation principles adopted by similar languages, mostly relating to comparative constructions, emphasizing adverbials, sentence-initial discourse phenomena and expletives.

For example, the Slovenian guidelines for the *expl* relation, which was previously used for labelling all instances of the reflexive pronouns *si* and *se* '(to) oneself', have now been improved so as to distinguish between true expletives (e.g. reflexive clitics as part of inherently reflexive verbs or passive constructions) and pronouns occurring as objects (Figure 1).

The official Slovenian UD guidelines are freely available both in Slovenian (as a standalone document)[3] and English (as part of the official UD website).[4] In addition to the category-based description of the universal guidelines and its application to specific examples in Slovenian, the guidelines document also features a construction-based appendix, in which the treatment of specific syntactic phenomena is addressed, including the challenging constructions identified within the annotation campaign described in Section 5.

4. Revision of the SSJ Treebank

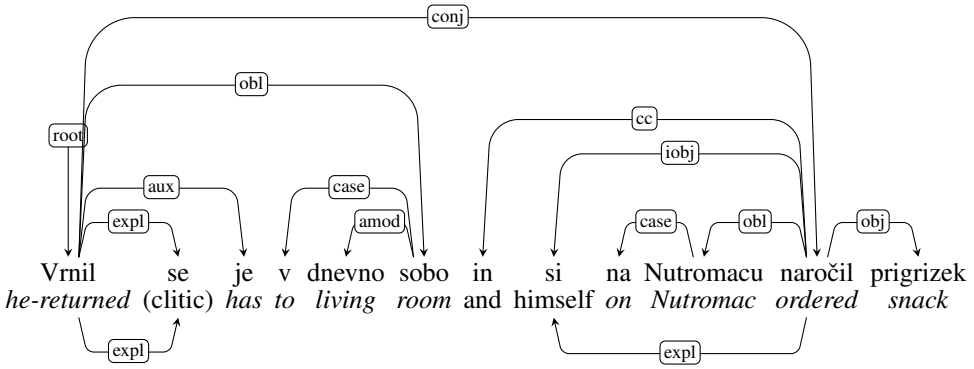
In the data preparation stage, the original SSJ treebank annotations were manually improved to implement the newly proposed changes in the annotation guidelines (Section 3) and remove the previously identified annotation mistakes and inconsistencies arising from the original conversion (Section 2). Among others, these included conflicting annotations of paratactical and coordinating clauses, direct and indirect objects, appositional structures, specific multi-word expressions, and a relatively high number of unjustified non-projective relations.

For each of the approximately 30 identified types of issues, various heuristics were used to identify sentences with potentially problematic annotations, which were then manually inspected and corrected in accordance with the guidelines. In the process, 1,670 relations in the original SSJ dataset have been corrected, the distribution of which reflects the structures mentioned above, as two thirds of the corrections pertained to the *advmod*, *nmod*, *obl*, *parataxis*, *appos* and *expl*

[2] The rules and conversion scripts from JOS to UD are available at `https://github.com/clarinsi/jos2ud`.

[3] The document will be published in accordance with the DSDE project timeline as part of the official project website. The preliminary version is available at `http://tiny.cc/ud-sl_guidelines`

[4] `https://universaldependencies.org/`



(He returned to the living room and ordered himself a snack from Nutromac.)

Figure 1: Example sentence from SSJ illustrating the change of Slovenian UD guidelines for the expletive *expl* relation from the initial treebank release (below) to UD release v2.10 (above).

relations. This manual work resulted in the slightly revised version of the original SSJ treebank, which was used as the basis for subsequent new data addition described in the sections below.

5. Extension of the SSJ Treebank

The extension of the old SSJ treebank was performed in two subsequent stages, in which new sentences from the ssj500k corpus were added and a new ELEXIS subset was created. In both stages, the data annotation was performed using the ssj500k-compliant Q-CAT corpus annotation tool (Brank, 2022), which was upgraded to also support the CONLL-U format, while the curation stage was performed using the WebAnno (Eckart de Castilho et al., 2016) web-based tool hosted by CLARIN.SI.[5]

5.1. Extension 1: Semi-Converted ssj500k

In the first stage of the project, the 3,411 sentences from the original ssj500k corpus which had not been fully converted from JOS to UD dependency trees at the time of original SSJ compilation (see Section 2) were manually inspected so that the tokens with missing (unconverted) UD dependency annotations were also labeled. Specifically, the semi-converted dataset included 95,194 tokens, out of which 22,377 tokens (23.5%) were initially labeled as *unknown* dependents of the root node (Figure 2). This means that on average 6.6 dependency relations per sentence had to be manually created from scratch, while the existing relations were also checked for the accuracy of conversion.

The process was designed as a multi-annotator annotation campaign, in which each sentence was annotated by two independent annotators (pre-trained linguists) and the final curator in case of disagreements. Although it is difficult to report on the inter-annotator agreement given the specificity of the task (manual corrections of partially converted data), on average, the two annotators agreed on 92.1% annotations (87,675 out of 95,194 tokens). For the *unknown* relations in particular, the absolute agreement was much lower (80.5% or 18,023 out of 22,377 tokens), but it was expected given the complexity of the task (annotation of the most complex syntactic constructions in long sentences).

In total, the activity resulted in 22,377 newly added dependency relations and 4,623 corrected dependency relations in the semi-converted ssj500k subset, amounting to 27,000 (28.4%) tokens with corrected annotations. Almost one half of the previously unlabeled (*unknown*) were punctuation tokens (*punct*), which was expected given the original mapping rules, where punctuation attachment was performed after most other sentence annotations were known. This includes the identification of the sentence *root* element, which is the second most frequent type of unconverted tokens (12%), followed by *parataxis* (9%) and (mostly clausal) coordination (*conj*, 6%), confirming the type of constructions reported to be the most challenging at the time of the original ssj500k conversion to SSJ (Section 2).

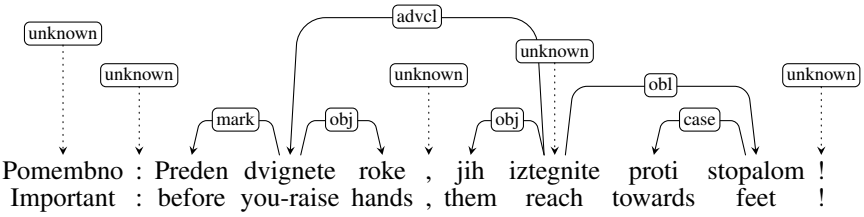
On the other hand, most corrections of successfully converted labels pertained to change of head attachment for adverbial modifiers (*advmod*, 20% of all corrections) and punctuation (16%), while most label corrections involved the switch from nominal modifiers (*nmod*) to prepositional adjuncts (*obl*, 4%), with other corrections being more equally distributed across individual relations. This, however, does not necessarily reflect the accuracy of the original rule-based conversion, given the semi-converted sentences do not reflect the final output for the fully converted sentences, as illustrated by the inter-dependent rules for punctuation attachment in the paragraph above.

Given the long learning curve related to this relatively complex annotation task, the speed of the annotators varied from an average of 11 sentences (307 tokens) per hour in the beginning of the first stage to approximately 15 sentences (419 tokens) per hour at its completion.

5.2. Extension 2: ELEXIS

In the second stage of the project, the second new SSJ subset was created based on the ELEXIS-WSD-

[5] https://www.clarin.si/webanno/.



(Important: Before raising your hands, reach them towards your feet.)

Figure 2: An example of a semi-converted ssj500k sentence with some missing (*unknown*) dependency annotations.

SL corpus, the Slovenian subset of the ELEXIS parallel sense-annotated dataset (Martelli et al., 2021) extracted from WikiMatrix (Schwenk et al., 2021), a large open-access collection of (translated) parallel data derived from Wikipedia. The corpus comprised of 2,024 Slovenian sentences (31,237 tokens) with manual annotations of tokenization, lemmatization and JOS morphosyntactic annotations.

For UD morphology (POS tags, morphological features), the existing mapping scripts (Section 2) were used for conversion from JOS to UD, followed by a manual disambiguation of the AUX and VERB instances of the verb *biti*. Afterward, the dataset was parsed with UD dependency relations using the classla-stanza parsing tool (Ljubešić and Dobrovoljc, 2019) trained on the concatenation of the available data, i.e. the slightly revised original SSJ treebank (Section 4) and the new ssj500k-based extension (Section 5.1).

The automatically parsed ELEXIS dataset was then manually checked by three annotators and the final curator. In the process, 1,534 dependency relations have been manually corrected (854 for wrong head, 252 for wrong relation and 428 for both), mostly pertaining to constructions labeled as *nmod*, *advmod*, *obl*, *conj* and *punct*. The indirectly observed parsing accuracy (95%) was in line with the expected parser performance on standard written texts (see, for example, evaluations reported in Table 3), and was also reflected in the annotation speed (an average of 37.5 sentences per hour) and a relatively high inter-annotator agreement (96.3% identically annotated tokens by the three annotators).

5.3. Overview of the New SSJ Treebank

Finally, the slightly revised original treebank (Section 4) and the two newly available datasets described in Sections 5.1 and 5.2 have been merged into the new–improved and extended–version of the Slovenian SSJ treebank (Table 1), which has been released in UD v2.10.[6]

As shown in Table 1, the SSJ treebank size has increased by 5,435 sentences (+67.9%) and 126,427 tokens (+89.9%) in comparison to the old version and now places as the 30th out of 218 UD treebanks ranked according to the number of words. In the continuation of the paper, we evaluate and discuss the impact of this substantial data increase on the state-of-the-art dependency parsing of Slovenian.

Subset	Sent.	Tokens	Avg.len.
Old SSJ	8,000	140,670	17.58
Ext. 1 (ssj500k)	3,411	95,194	27.91
Ext. 2 (ELEXIS)	2,024	31,233	15.43
Total	13,435	267,097	19.88

Table 1: Overview of the new version of Slovenian SSJ UD treebank.

6. Evaluation

We evaluate the extensions of the SSJ Universal Dependencies treebank by training the Bi-LSTM based biaffine parser (Dozat and Manning, 2016) available in the classla-stanza pipeline (Ljubešić and Dobrovoljc, 2019), a fork of the Stanza pipeline (Qi et al., 2020), the fork containing many improvements primarily on the level of tokenization, part-of-speech tagging and lemmatization.

Besides the parsing training data, we use the CLARIN.SI-embed.sl embedding collection (Ljubešić and Erjavec, 2018) trained on a 3.5 billion tokens collection of Slovenian texts.

In the following subsections we address the decisions of the new SSJ data splits in light of the newly available data, our experimental setup, and, finally, the results of our experiments.

6.1. Data splits

While constructing the new data split, a prerequisite for the official data release, we mostly followed the decisions from the old SSJ dataset, which was a consecutive split, i.e., the first part of the dataset was used as a training dataset, with the two latter parts being the dev and the test datasets, in a rough 8:1:1 sentence-based ratio. Given that Extension 1 of the SSJ dataset, coming from the same text source as the original SSJ (ssj500k), consists of sentences roughly uniformly distributed across the whole dataset, it was possible to keep a similar divide of the newly added ssj500k data, with minor deviations due to the fact that in the new split document

[6]This paper reports work on a penultimate release version (commit fa057316b79779893659bdc007cdc7f6465e58f3), which has been slightly changed for the official UD v2.10 release due to stricter validation rules. Namely, approximately 50 annotations were changed, equally distributed across various morphological and syntactic categories.

boundaries were respected, which was previously not the case. The large majority of the original SSJ sentences was thus preserved in the same (data, dev or test) portion of the dataset.
Given the specificities of each of the two data extensions (the first extension adding significantly longer and syntactically more complex sentences, the second extension adding shorter translated sentences from Wikipedia), Extension 2 data was added to all three data splits as well, again in a 8:1:1 ratio. This decision also allows for a more diverse evaluation, which will prove to be rather useful for identifying potential biases in our data.
The final size of the split, measured in number of sentences, is given in Table 2. The split shows for all three splits of the extended SSJ dataset to consist of 85% of sentences from the ssj500k dataset and the remainder from the ELEXIS dataset.

Subset	Train	Dev	Test
Old SSJ	6,478	734	788
Extension 1	2,807	313	291
Extension 2	1,618	203	203
New SSJ	10,903	1,250	1,282

Table 2: Comparison (in number of sentences) of the old SSJ dataset split and the new SSJ dataset split, by each data extension.

6.2. Experimental setup

To perform an automatic investigation of the potential gains obtained by the SSJ data extension described in Section 5 in comparison to the original SSJ data, we have trained two parsers:

1. The SSJ_old model based on the original SSJ dataset (as described in Section 4)
2. The SSJ_new model based on the original SSJ dataset with both extensions, from the ssj500k (5.1) and the ELEXIS datasets (Section 5.2).

Both during training, development and testing, we use gold-annotated upstream data (tokenization, sentence splitting, morphosyntactic tagging, lemmatization) as we are primarily interested in the improvements in the dependency parsing performance, and not the interplay of automatic upstream and dependency syntax annotations. Using automatically annotated upstream layers with different taggers and lemmatizers (based on the specific training data) would have blurred the impact our data interventions had on the dependency syntax layer, especially given that classla-stanza tagging and lemmatization depend also on additional external processes and data such as rule-based tokenizers that partially perform tagging, inflectional lexicons, and additional training data.
We evaluate both parsers on the new SSJ test set (Table 2, 1,282 sentences) using the standard label attachment score (LAS) that gives the percentage of nodes with correctly assigned parent node and the type of relation between them. To explore the scalability of the new model to different data types, as well as the potential differences in the three main SSJ subsets, the evaluation results are also reported for each test subset individually, i.e. the old SSJ test data (788 sentences), the Extension 1 test data (291 sentences) and the Extension 2 test data (203 sentences).
We report both on the overall parsing performance (Section 6.3) and the performance for individual relations (Section 6.4).

6.3. Overall Performance

The overall parsing performance, reported in Table 3, confirms the general benefits of our data extensions, with the improvement of 1.85 LAS on the whole new SSJ test set (21% relative error reduction).[7] As expected, the biggest increases are observed for the two newly added subsets, that is +4.41 LAS (29% error reduction) for Ext-1 ssj500k data and +2.11 LAS (39% relative error reduction) for Ext-2 ELEXIS data, while the benefits of the new model are much less pronounced when evaluated on the old SSJ data alone (+0.3, 5.5% relative error reduction).

Models	Test datasets			
	New SSJ	Old	Ext-1	Ext-2
SSJ_old	91.36	94.53	84.67	94.60
SSJ_new	93.21	94.83	89.09	96.71

Table 3: Results of the automatic evaluation of the old and new SSJ model, measured through labeled attachment score (LAS) F1 on the new SSJ test set and its subsets.

This confirms the distinct nature of the original and the two newly added datasets, which, as already reported above, differ in terms of data source, sentence length, tree complexity and source of annotations. The higher parsing scores on the old SSJ and the Ext-2 ELEXIS test data in Table 3 for both models suggest that the old SSJ and the Ext-2 ELEXIS data are more similar and easier to parse in general, which is line with the shorter average sentence length reported in Table 1. On the other hand, the newly added Ext-1 ssj500k data is definitely the hardest test set of the three, which is expected given it mostly includes sentences that were too complex to be covered by the automatic rule-based conversion at the time of the original SSJ treebank creation (Section 2).
The original SSJ dataset was therefore potentially biased towards simpler sentences, which is not only illustrated by the seeming drop in performance when com-

[7] We calculate the relative error reduction as the percentage of the difference of the LAS score between the new and the old system in the difference between a perfect LAS score and the old LAS score, i.e. $(LAS_{new} - LAS_{old})/(100 - LAS_{old}) * 100$.

paring the evaluation of the old model on the old test set (94.53 LAS) and the new model on the new test set (93.21 LAS), but has also been suggested by the results of the CoNLL 2018 Shared Task (Zeman et al., 2018)[8] and the official Stanza evaluations,[9] which list the (old) SSJ treebank as one of the highest-ranking treebanks according to LAS score.

6.4. Relation-Based Performance

To better understand which specific constructions benefit from the newly available data and what can be expected from the new model when used for specific parsing tasks in downstream applications, we extend the overall evaluation on the new SSJ test set described in Section 6.3 above to individual dependency relations as well.

As shown in Table 4,[10] with the exception of *vocative* and *dep*,[11] all relations demonstrate an increase of LAS F1 in comparison to the baseline old SSJ. Specifically, the biggest improvements are gained for relations pertaining to constructions which were rare or under-represented in the old SSJ treebank, but are frequent in the newly added data (ssj500k in particular), such as lists (*list*, +75.86 F1), elliptical structures (*orphan*, +68.24 F1), appositional modifiers (*appos*, +13.40 F1) and discourse particles (*discourse*, +9.97 F1), with a significant error reduction also observed for fixed multi-word expressions (*fixed*, 52%) and numeric modifiers (*nummod*, 43%).

On the other hand, the smallest gains are observed for adjectival modifiers (*amod*, +0.10pp, 8% relative error reduction), clausal complements (*ccomp*, +0.23pp, 2% relative error reduction) and expletives (*expl*, +0.40pp, 11% relative error reduction), suggesting that increasing the size of the training data and making it more diverse is less significant for some of the relations.

In absolute terms, the new model is most successful in parsing function words, such as prepositions (*case)*, auxiliary verbs (*aux*), determiners (*det*) and subordinating conjunctions (*mark*), as well as adjectival modifiers of nominals (*amod*), which all exhibit LAS above 98 F1. For core semantic phenomena, which are typically the most relevant relations for various downstream applications, above average performance is observed for nominal subjects (*nsubj*) and objects (*obj*). On the other hand, indirect objects (*iobj*, adjuncts (*obl, advmod*) and their clausal counterparts (*ccomp, csubj, advcl*) still exhibit below-average performance despite some important improvements based on the newly available data (e.g. +5.80 LAS and 29% error reduction for *csubj* clausal subjects).

7. Conclusion

We have presented a new version of the reference SSJ Universal Dependencies Treebank for Slovenian, which has been revised and extended to almost double the original size. The process was based on the initial revision and exhaustive documentation of the language-specific UD annotation guidelines for Slovenian, followed by a systematic multi-stage annotation campaign, in which the original SSJ data has been slightly revised and substantially extended by new sentences coming from the ssj500k and ELEXIS corpora.

After proposing the official UD data splits for the extended SSJ treebank, the data was used to train a new dependency parsing model in the classla-stanza NLP tool,[12] and compare its performance to the model trained on the old, un-extended SSJ data.

At first glance, the results seem very unimpressive given the labour-intensive data annotation campaign, with only marginal performance gains when the two models are evaluated on the old test data. However, when evaluated on the new, extended and diversified data, the parsing performance improvements are substantially more pronounced, especially for previously under-represented syntactic phenomena, which have mostly been left out of the original SSJ due to the limitations of its rule-based creation.

The new diversified SSJ dataset might therefore introduce some new challenges to the parsing systems, but will also make them much more accurate with respect to naturally occurring language data. This is especially important for under-resourced languages like Slovenian, where large-scale development of domain-specific treebanks and parsing systems cannot be realistically expected.

Nevertheless, given the now empirically confirmed distinct nature of the three SSJ subsets, future work should be dedicated to a systematic in-depth investigation of the possible points of divergence between the datasets with respect to parsing performance, such as text source, sentence length, tree complexity and possible annotation inconsistencies, which could potentially lead to new insights for further SSJ treebank consolidation and extension on the one hand, and parsing system modifications on the other.

Last but not least, although our evaluation was deliberately focused on dependency parsing only, the new SSJ dataset represents an equally important contribution to the development of lemmatization, part-of-speech tagging and other models for morphological processing of

[8] https://universaldependencies.org/conll18/results-treebanks-las.html

[9] https://stanfordnlp.github.io/stanza/performance.html

[10] Relations not occurring in the test set (*dislocated*, *goeswith*, *reparandum*) are excluded, while extensions (e.g. *flat:name*) are truncated to their universal counterparts (e.g. *flat*).

[11] The zero increase of performance for *vocative* and *dep* is expected, given that the former only has a single occurrence in the test set, while the latter is used for labelling irregular, marginal phenomena.

[12] The new parsing model is planned to be released on the CLARIN.SI repository and integrated into the classla-stanza pipeline: https://github.com/clarinsi/classla.

Relation	Description	SSJ_old	SSJ_new	F1 diff	RER
acl	clausal modifier of noun	80.76	81.73	0.97	5%
advcl	adverbial clause modifier	71.37	75.86	4.49	16%
advmod	adverbial modifier	87.01	89.95	2.94	23%
amod	adjectival modifier	98.8	98.9	0.1	8%
appos	appositional modifier	50	63.4	13.4	27%
aux	auxiliary verb	98.45	98.93	0.48	31%
case	case marking preposition	98.72	99.17	0.45	35%
cc	coordinating conjunction	94.94	96.27	1.33	26%
ccomp	clausal complement	90.44	90.67	0.23	2%
conj	conjunct	81.52	85.91	4.39	24%
cop	copula verb	93.4	95.43	2.03	31%
csubj	clausal subject	79.73	85.53	5.8	29%
dep	unspecified dependency	54.55	54.55	0	0%
det	determiner	98.31	98.79	0.48	28%
discourse	discourse element	59.26	69.23	9.97	25%
expl	expletive	96.31	96.71	0.4	11%
fixed	fixed multi-word expression	86.03	93.33	7.3	52%
flat	flat multi word-expression	87.97	92.12	4.15	35%
iobj	indirect object	78.57	81.66	3.09	14%
list	list	0	75.86	75.86	76%
mark	marker (subordinating conjunction)	97.88	98.69	0.81	38%
nmod	nominal modifier	85.72	87.44	1.72	12%
nsubj	nominal subject	93.69	95.28	1.59	25%
nummod	numeric modifier	89.95	94.23	4.28	43%
obj	(direct) object	95.08	95.53	0.45	9%
obl	oblique nominal (adjunct)	89.59	91.14	1.55	15%
orphan	dependent of missing parent	0	68.24	68.24	68%
parataxis	parataxis	63.32	70.35	7.03	19%
punct	punctuation symbol	90.3	93.08	2.78	29%
root	root element	95.09	96.26	1.17	24%
vocative	vocative	0	0	0	0%
xcomp	open clausal complement	91.71	92.87	1.16	14%
ALL	all relations	91.36	93.21	1.85	21%

Table 4: Relation-based comparison of LAS F1 performance of the parsing model trained on the old and the new SSJ data with relations listed alphabetically. The last two columns give the absolute F1 difference and the relative error reduction (RER).

Slovenian, especially for systems trained on UD treebanks alone.

8. Acknowledgements

The work described in this paper was supported by the project Development of Slovene in a Digital Environment co-financed by the Republic of Slovenia and the European Union from the European Regional Development Fund, and the research program "Language Resources and Technologies for Slovene" (P6-0411) funded by the Slovene Research Agency. A special acknowledgment goes to the data annotation team (Tina Munda, Ina Poteko, Rebeka Roblek, Luka Terčon and Karolina Zgaga).

9. Bibliographical References

Arhar Holdt, Š. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2).

Brank, J. (2022). Q-CAT corpus annotation tool 1.3. Slovenian language resource repository CLARIN.SI.

De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational linguistics*, 47(2):255–308.

Dobrovoljc, K., Erjavec, T., and Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*, pages 33–38.

Dobrovoljc, K., Erjavec, T., and Ljubešić, N. (2019). Improving UD processing via satellite resources for morphology. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 24–34, Paris, France, August. Association for Computational Linguistics.

Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., and Dobrovoljc, K. (2020a). Gigafida 2.0: The reference corpus of written standard Slovene. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3340–3345, Marseille, France, May. European Language Resources Association.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, , Čibej, J., and Brank, J. (2020b). The ssj500k training corpus for Slovene language processing. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 24–33, Ljubljana, Slovenia, September. Institute of Contemporary History.

Ljubešić, N. and Dobrovoljc, K. (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*, pages 29–34, Florence, Italy, August. Association for Computational Linguistics.

Martelli, F., Navigli, R., Krek, S., Tiberius, C., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Pedersen, B., Olsen, S., Langements, M., Koppel, K., Üksik, T., Dobrovolijc, K., Ureña-Ruiz, R.-J., Sancho-Sánchez, J.-L., Lipp, V., Varadi, T., Györffy, A., László, S., Quochi, V., Monachini, M., Frontini, F., Tempelaars, R., Costa, R., Salgado, A., Čibej, J., and Munda, T. (2021). Designing the ELEXIS parallel sense-annotated dataset in 10 european languages. In *eLex 2021 Proceedings*, eLex Conference. Proceedings. Lexical Computing CZ. null ; Conference date: 05-07-2021 Through 07-07-2021.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, April. Association for Computational Linguistics.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (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, October. Association for Computational Linguistics.

10. Language Resource References

Ljubešić, Nikola and Erjavec, Tomaž. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0*.

Zeman, Daniel and others. (2022). *Universal Dependencies 2.10*.