Doubling the Amount of Training Data: Does It Help? A New Training Corpus for Slovene and Its Impact on Automatic UD Annotation

Luka Terčon
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, 1000 Ljubljana
luka.tercon@fri.uni-lj.si

Nikola Ljubešić
Jožef Stefan Instute
Jamova cesta 39, 1000 Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Kaja Dobrovoljc
Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
kaja.dobrovoljc@ff.uni-lj.si

Relevant UniDive working groups: WG1, WG3

1 Introduction

The ssj500k training corpus was until recently the largest collection of manually annotated training data for Slovene (Krek et al., 2020), containing about 500,000 tokens, annotated on various levels of linguistic annotation. As part of the ongoing Development of Slovene in a Digital Environment project (Slovene: Razvoj slovenščine v digitalnem okolju - RSDO),[1] we expanded this corpus with approximately 500,000 more tokens, thus reaching the figure of about 1 million tokens in total. The resulting dataset was named the Slovene Training Corpus (Slovene: Slovenski učni korpus) or the SUK corpus and was published in December 2022 via the CLARIN.SI repository (Špela Arhar Holdt et al., 2022).

The aim of this abstract is to present the structure of the new training data and describe how it was used to train the CLASSLA-Stanza tool for automatic linguistic annotation (Ljubešić and Dobrovoljc, 2019).[2] We specifically focus on investigating how varying the amount of training data impacts the performance of the tool for universal part-of-speech tag and universal dependency relation prediction. The impact of adding an inflectional lexicon to the classifier tool as a controlling element is also explored.

[1] https://slovenscina.eu/
[2] https://pypi.org/project/classla/

2 The new training corpus

The SUK corpus consists of 2,908 documents split up into 11,516 paragraphs and 48,594 sentences, all together amounting to 1,025,639 tokens. Much like its predecessor, it is composed of a balanced selection of text genres, covering both fiction and non-fiction. In addition to ssj500k—its core part—it also includes three new parts which were manually annotated in recent annotation campaigns: the Slovene part of the Parallel sense-annotated corpus ELEXIS-WSD 1.0 (Martelli et al., 2022), the Slovene corpus for aspect-based sentiment analysis SentiCoref 1.0 (Žitnik, 2019), and Ambiga—a corpus containing a number of ambiguous word forms constructed with the aim of improving model performance on ambiguous instances.

The SUK corpus contains annotations on several levels of grammatical description. The entire corpus is annotated on the level of sentence segmentation, lemmatization, and morphosyntactic tagging, while other annotations have only been applied to parts of the whole. The numbers associated with each annotation layer are shown in Table 1.

3 Automatic annotation experiments

Our experiments focused on three layers of automatic grammatical annotation: lemmatization, part-of-speech tagging, and syntactic dependency parsing (both using the Universal Dependencies[3] framework for grammatical annotation). We tested

[3] https://universaldependencies.org/

Annotation layer	ssj500k	SUK
Segmentation	586,248	1,025,639
Lemmatization	586,248	1,025,639
JOS morphosyntax	586,248	1,025,639
UD PoS and features	586,248	1,025,639
UD dependencies	140,670	267,097
JOS dependencies	235,864	267,097
Semantic roles	112,048	209,791
Named entities	194,637	659,059
Verbal MWEs	280,522	280,522
Coreference chains	n/a	391,962

Table 1: Number of tokens annotated in the old ssj500k and the new SUK training corpus on each layer of annotation.

how much of an effect two variations in the training and evaluation procedure have on the final annotation tool performance scores.

For the first variation, the amount of training data was doubled, for the second, we added an inflectional lexicon to the annotation tool as a controlling element for the predictions.

The CLASSLA-Stanza tool for language processing, which includes lemmatization, POS tagging, and dependency parsing models for Slovene, was used to perform the experiments. Several combinations of linguistic models were produced by training on different amounts of training data.

For the dependency parsing annotation layer, every model was trained only on the subset of the corpus that contains UD syntactic dependency annotations (267,097 tokens for **SUK** and 140,670 tokens for **ssj500k**.).

The following metrics were used to evaluate model performance on each annotation layer: **POS tagger** - F1 score for all morphosyntactic tags (XPOS, UPOS and UFEATS), **Lemmatizer** - F1 score for all lemmas, **Dependency parser** - F1 of the labeled attachment score (as defined in Zeman et al. 2018).

3.1 Doubling the amount of training data

In the first experiment, a first set of models was trained on the ssj500k training data and a second set on the new SUK training data. This way we investigated how models trained on the original training set compare to models trained on double the amount of training data. The models' performance before and after the doubling is displayed in Table 2. The results show a clear improvement in performance after doubling the amount of training data.

Annotation layer	Dataset	Score
POS tagger	ssj500k	96.61
	SUK	**97.55**
Lemmatizer	ssj500k	98.89
	SUK	**99.33**
Dependency parser	ssj500k	87.78
	SUK	**91.06**

Table 2: Comparison of model performance before and after doubling the amount of training data. Bold results are statistically significantly different to the alternative.

3.2 Adding an inflectional lexicon into the mix

For the second experiment, we analyzed model performance before and after adding an inflectional lexicon as a controlling element. This method restricts the model predictions to match the forms and combinations present within the lexicon. For inflectionally rich languages such as Slovene, it has been shown that this approach can improve model performance on certain tasks, especially with large training corpora (Ljubešić and Erjavec, 2016; Ljubešić and Dobrovoljc, 2019). The Sloleks morphological lexicon for Slovene (Čibej et al., 2022)—more than 300,000 entries in size—was used in our experiments.

The results are shown in Table 3. The lexicon clearly improves POS tagger performance, however on the level of lemmatization and dependency parsing the results are much more difficult to interpret. A subsequent error analysis showed that in some instances the lexicon guidance did improve the results, but also that the lexicon we used contains a number of automatically-generated entries, which proved detrimental to the performance of the annotation tool in some instances.

4 Conclusion

The experiments presented demonstrate that the increased amount of training data present in the new training corpus for Slovene improves the performance of tools for automatic grammatical annotation. This trend holds for all three inspected annotation layers. However, introducing an inflectional lexicon to limit the model predictions does not lead to a consistent improvement in the performance scores except for morphosyntactic tagging. Rather, it may lower the accuracy of the predictions, due

Annotation layer	Dataset	Lexicon usage	Score
POS tagger	ssj500k	yes	**96.98**
		no	96.61
	SUK	yes	**97.94**
		no	97.55
Lemmatizer	ssj500k	yes	98.68
		no	**98.89**
	SUK	yes	99.11
		no	**99.33**
Dependency parser	ssj500k	yes	87.67
		no	87.78
	SUK	yes	91.11
		no	91.06

Table 3: Comparison of model performance with and without lexicon usage. Bold results are statistically significantly different to the alternative.

to some problematic entries in the lexicon. This outcome reflects the great importance of good quality manually annotated data when it comes to both training corpora and inflectional lexicons.

Acknowledgements

The work described by this paper was made possible by the Development of Slovene in a Digital Environment project (Razvoj slovenščine v digitalnem okolju), financed by the Ministry of Culture of the Republic of Slovenia and the European Regional Development Fund, and the Language Resources and Technologies for Slovene research program (P6-0411), financed by the Slovenian Research Agency from the state budget. We also thank the annotators for annotating the new data (Tina Munda, Ina Poteko, Rebeka Roblek, Luka Terčon, Karolina Zgaga) and Tomaž Erjavec, Luka Krsnik, Cyprian Laskowski and Mihael Šinkec for technical support.

References

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Janez Brank. 2020. The ssj500k Training Corpus for Slovene Language Processing. pages 24–33.

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'16)*, pages 1527–1531, Portorož, Slovenia. European Language Resources Association (ELRA).

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. pages 29–34. Association for Computational Linguistics.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, and Tina Munda. 2022. Parallel sense-annotated corpus ELEXIS-WSD 1.0. Slovenian language resource repository CLARIN.SI.

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. pages 1–21. Association for Computational Linguistics.

Jaka Čibej, Kaja Gantar, Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Luka Krsnik, and Marko Robnik-Šikonja. 2022. Morphological lexicon Sloleks 3.0. Slovenian language resource repository CLARIN.SI.

Špela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2022. Training corpus SUK 1.0. Slovenian language resource repository CLARIN.SI.

Slavko Žitnik. 2019. Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0. Slovenian language resource repository CLARIN.SI.