Software Impacts 8 (2021) 100063

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

OldSlavNet: A scalable Early Slavic dependency parser trained on modern language data

Nilo Pedrazzini *, Hanne Martine Eckhoff

University of Oxford, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:
Neural networks
Dependency parsing
Cross-lingual transfer
Treebanks
Early Slavic

ABSTRACT

Historical languages are increasingly being modelled computationally. Syntactically annotated texts are often a sine-qua-non in their modelling, but parsing of pre-modern language varieties faces great data sparsity, intensified by high levels of orthographic variation. In this paper we present a good-quality Early Slavic dependency parser, attained via manipulation of modern Slavic data to resemble the orthography and morphosyntax of pre-modern varieties. The tool can be deployed to expand historical treebanks, which are crucial for data collection and quantification, and beneficial to downstream NLP tasks and historical text mining.

Code metadata

Current code version	Rolling release (currently v2.0)
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2021-7
Permanent link to Reproducible Capsule	https://codeocean.com/capsule/8481687/tree/v1
Legal Code Licence	MIT Licence
Code versioning system used	Git
Software code languages, tools, and services used	Python, Ruby
Compilation requirements, operating environments & dependencies	Python 3.x, DyNet 2.0; see also https://github.com/npedrazzini/OldSlavNet/blob/master/requirements.txt
If available Link to developer documentation/manual	https://npedrazzini.github.io/OldSlavNet
Support email for questions	nilo.pedrazzini@ling-phil.ox.ac.uk

1. Introduction

Dependency parsing is important in many downstream natural language processing (NLP) tasks, including event extraction, word vector representation enhancement, and text classification and summarization. Training good-quality parsers for historical languages is a challenging task, since they normally provide very little data with very high levels of linguistic variation, which in machine learning easily translates into high levels of noise.

In this paper we present a variety-agnostic part-of-speech (PoS) tagger and dependency parser for Early Slavic (OldSlavNet) trained on multi-lingual Slavic data spanning a thousand years via orthographic and morphosyntactic harmonization of the modern data with their pre-modern counterparts. Early Slavic and Modern Russian data was obtained from the Tromsø Old Church Slavonic and Old Russian Treebanks (TOROT) [1] (specifically the entirety of its Church Slavonic and Old Russian subcorpus, and part of SynTagRus [2] from its Modern Russian one), whereas Modern Serbian data was collected from the Universal Dependencies (UD) Serbian-SET treebank [3]. Unlike other experiments on historical language parsing (e.g. [4] on pre-modern Germanic), and similar transfer techniques for the morphological tagging of historical languages, including Slavic (e.g. [5]), OldSlavNet does not rely on annotation projection between historical texts and their modern translations, but on modern data from a variety of contemporary resources of different genres. This is a particularly welcome feature, since it makes the parser a scalable tool: additional training data can be added from original contemporary sources of any genre to improve the parser without training it from scratch. The harmonization scripts,

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail address: nilo.pedrazzini@ling-phil.ox.ac.uk (N. Pedrazzini).

https://doi.org/10.1016/j.simpa.2021.100063
Received 23 January 2021; Received in revised form 11 February 2021; Accepted 12 February 2021

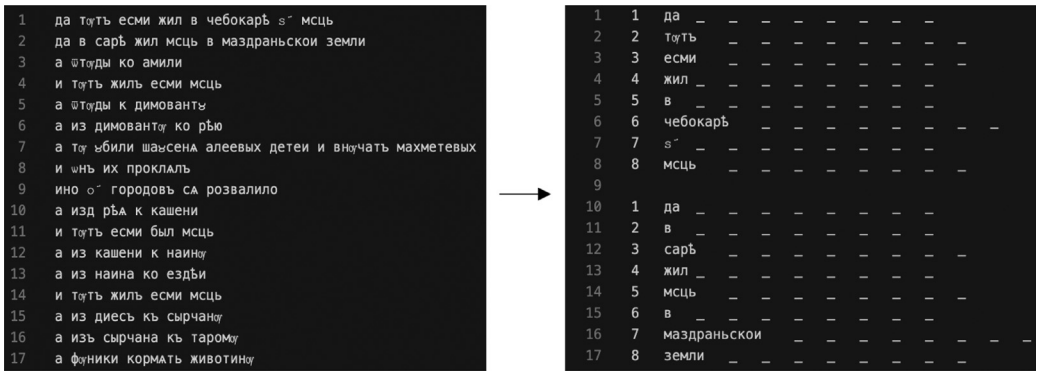

Fig. 1. Example of a tokenized, one-sentence-per-line text file (left) and its CoNLL-U version (right) obtained using `converter.py`.

now available for Russian and Serbian, can be downloaded from the parser's repository and used to harmonize new Modern Russian and Serbian texts with Early Slavic, thus potentially improving the parsing performance.

The parser is especially crucial to expand historical treebanks, large collections of digital texts annotated with syntactic information: treebanks are a versatile source of data, not only directly exploited in many NLP tasks, as the aforementioned ones, but they are used by the humanities at large as a stand-alone collection of carefully digitized textual data enriched with linguistic information.

2. Data and parser architecture

The parser works in the UD framework [6], one of the most widely employed formats for dependency parsing.

The tool's neural-network architecture is based on jPTDP [7]. The following are the main new features in OldSlavNet's model:

- `ArgParse` substitutes the older `OptParse` to allow for wider reusability of our code.
- RMSProp [8] is employed instead of Adam [9] as optimizer to avoid exploding gradients. The initial learning rate was set to `0.1` instead of `None`.
- Since the previous experiment in [10], the training set has been expanded with Modern Russian and Serbian data. OldSlavNet's documentation contains a detailed breakdown of the corpus on which the parser was trained and tested.

3. Usage

The following is the end-to-end process to use the tool to tag new Early Slavic text:

1. *Pre-process your text file*: Convert your Early Slavic text to the CoNLL-U UD-format by running the `converter.py` script included in OldSlavNet's repository. The input must be an already tokenized, one-sentence-per-line text file. Fig. 1 shows how typical input text files and output CoNLL-U files should look like.
2. *Download the repository*: Clone OldSlavNet's repository or download the relevant files:
   a. The Python scripts (`parser.py`, `oldslavdep.py`, `learner.py`, `decoder.py`, `mnnl.py` and `utils.py`).
   b. the `model` and `model.params` files.
   c. the `requirement.txt` file.
3. *Install the required dependencies*:
   Run: `pip install -r requirements.txt`
4. *Tag your CoNLL-U file*: Run the `parser.py` script and the necessary hyperparameters as detailed in the relevant documentation's section. Your output should look as in Fig. 2

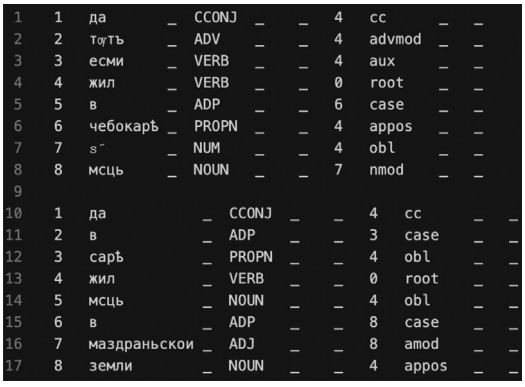

Fig. 2. Example of a CoNLLU file after adding part-of-speech and syntactic tags using OldSlavNet.

4. Impact

OldSlavNet's previous version (known as jPTDP-GEN) enabled [10], which discussed the improvement of dependency parsers for low-resource historical languages using cross-dialectal data. OldSlavNet, a generic (i.e. variety-agnostic) parser, was shown to perform better than two variety-specific parsers for Early Slavic, indicating that markedly non-standardized historical languages are likely to benefit more from the development of generic, cross-variety models, than from specialized ones. Since [10], OldSlavNet has further improved its real-world performance (i.e. its ability to tackle a wider range of pre-modern Slavic varieties and genres) thanks to additional data from Modern Russian and Modern Serbian, as Table A.1 shows.

OldSlavNet has been trialled on new texts in the TOROT Treebank [1,2], a major annotated historical corpus for Slavic and offspring of the PROIEL project [11,12]. The expansion of historical Slavic treebanks using OldSlavNet will contribute to the advancement of research domains that benefit from syntactically annotated data, particularly from less-resourced languages with great spelling variation:

1. *Semantic change detection*: A methodological gap which has been noted for decades [13] is the integration of syntactic information in meaning change modelling. Early Slavic treebank data can now be used in semantic change detection by generating word representation that are both semantically and syntactically constrained (e.g. syntactic word embeddings [14] and syntactic topic models [15]), thus improving the semantic models themselves. Understanding the mechanisms of meaning change in different historical contexts will help design better tools for semantic change detection, which has a wide range of applications in text processing, including information retrieval [16–18],

1	Oni	oni
2	su	biti
3	se	sebe
4	složili	složiti
5	da	da
6	pomognu	pomoći
7	iračkoj	irački
8	vladi	vlada
9	u	u
10	obuci	obuka
11	snaga	snaga
12	bezbednosti	bezbednost
13	i	i
14	ponudili	ponuditi
15	saradnju	saradnja
16	celokupnom	celokupan

1	Они	они
2	сѫть	бити
3	сѧ	себе
4	сложили	сложити
5	да	да
6	помогнѫть	помощи
7	ирачкои	ирачки
8	влади	влада
9	оу	оу
10	обоуци	обоука
11	снага	снага
12	безбедности	безбедност
13	и	и
14	поноудили	поноудити
15	сарадньѫ	сарадньа
16	целокоупномъ	целокоупан

Fig. 3. Example of Modern Serbian data before and after harmonization with Early Slavic orthography and morphology.

1	С	С
2	другой	ДРУГОЙ
3	стороны	СТОРОНА
4	вероятностный	ВЕРОЯТНОСТНЫЙ
5	алгоритм	АЛГОРИТМ
6	может	МОЧЬ.ipf
7	и	И
8	никогда	НИКОГДА
9	не	НЕ
10	выдать	ВЫДАВАТЬ.pf
11	результат	РЕЗУЛЬТАТ
12	но	НО
13	вероятность	ВЕРОЯТНОСТЬ
14	этого	ЭТО
15	равна	РАВНЫЙ
16	0	0

1	съ	С
2	другъıѣѧ	ДРУГОЙ
3	стороны	СТОРОНА
4	вероятностнъı	ВЕРОЯТНОСТНЫЙ
5	алгоритмь	АЛГОРИТМ
6	можетъ	МОЧЬ.ipf
7	и	И
8	никогда	НИКОГДА
9	не	НЕ
10	выдати	ВЫДАВАТЬ.pf
11	результатъ	РЕЗУЛЬТАТ
12	нъ	НО
13	вероятность	ВЕРОЯТНОСТЬ
14	этого	ЭТО
15	равна	РАВНЫЙ
16	0	0

Fig. 4. Example of Modern Russian data before and after harmonization with Early Slavic orthography and morphology.

culturomics [19], Diachronic Text Evaluation (DTE) [20,21], re-contextualization of past texts [22], OCR error correction [23], and abusive content detection [24], among others (see [25] for a detailed survey of applications).

2. *Improving NLP system evaluation practices*: Early Slavic is ideally placed to be used in the evaluation of NLP systems and methods, in light of its many related subvarieties and its high orthographic variation. This is a challenge in computational models of language change, since NLP systems tend to disregard low-frequency types, which are inevitable in historical sources. More syntactically annotated data for Early Slavic will allow us to systematically investigate how NLP approaches to infrequent tokens impact the generalization of a system's results, thus improving our evaluation practices.[1]
3. *Improving representativeness*: Expanding Early Slavic treebanks will allow us to develop methods for large-scale quantitative diachronic analyses of linguistic phenomena in languages other than English. The lack of large, non-English diachronic corpora has been stressed in the literature (e.g. [26] and [25]) as a possible bias in historical linguistic research that aims at generalizing findings cross-linguistically.

5. Limitations and future improvements

The scripts used to harmonize Russian and Serbian orthography and morphology to Early Slavic are still experimental. Presently, only the tokens belonging to the most frequent morphological tags have been harmonized. Figs. 3 and 4 illustrate how the harmonization routine currently works on a Serbian and a Russian sentence respectively. Given the promising results, in following releases we plan to develop harmonization scripts encompassing a wider range of morphotags, which is expected to yield even better parsing performance on pre-modern Slavic varieties.

[1] Note that the improvement of evaluation practices is a growing area of research in computational linguistics, with its dedicated venues, such as HumEval (https://humeval.github.io) and SemEval (https://semeval.github.io).

Table A.1
Performance of OldSlavNet compared to the previous state of the art.

Model	Test set	UAS	LAS
OldSlavNet	Codex Marianus	**84.12**	**78.92**
jPTDP-GEN [10]		83.79	78.42
OldSlavNet	Primary Chronicle (PVL)	85.33	79.66
jPTDP-ESL [10]		**85.70**	**80.16**
OldSlavNet	Vita Constantini	**70.72**	**56.64**
jPTDP-GEN		69.23	56.41
OldSlavNet	Codex Suprasliensis	**74.23**	**66.51**
jPTDP-GEN		72.28	63.38
OldSlavNet	Life of Sergij of Radonezh	**74.10**	**66.11**
jPTDP-GEN		73.90	65.76

A drawback of the current version of OldSlavNet is that it takes already sentencized text (i.e. with one sentence per line, as shown in Fig. 1) as an input, which requires users to manually split their text into sentences. Implementation of OldSlavNet with spaCy [27] is however underway, in order to complement the parser with an Early Slavic sentencizer that takes an unbroken texts as input and provides a one-sentence-per-line output, which can then be directly fed to OldSlavNet to add syntactic annotation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Economic and Social Research Council, United Kingdom of Great Britain and Northern Ireland [grant number ES/P000649/1].

We are thankful to Dat Quoc Nguyen for kindly giving their permission to experiment with the original neural-network model (jPTDP [7]) on which OldSlavNet has been developed.

Appendix. Parser performance

OldSlavNet has been developed to be applicable to different Early Slavic varieties and a wider range of genres than its previous, experimental version. In Table A.1 we report its performance on test sets which belong to various pre-modern Slavic dialects and present major orthographic and morphological differences.

References

[1] H.M. Eckhoff, A. Berdičevskis, Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank, Scr. e-Scripta 14–15 (2015) 9–25.
[2] A. Berdičevskis, H.M. Eckhoff, A Diachronic Treebank of Russian spanning more than a thousand years, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 5251–5256, https://www.aclweb.org/anthology/2020.lrec-1.646.
[3] T. Samardžić, M. Starović, Ž. Agić, N. Ljubešić, Universal Dependencies for Serbian in comparison with Croatian and other Slavic languages, in: Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 39–44, http://dx.doi.org/10.18653/v1/W17-1407.
[4] M. Sukhareva, C. Chiarcos, Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic, in: Proceedings of the First Workshop on Applying NLP Tools To Similar Languages, Varieties and Dialects, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014, pp. 11–20, http://dx.doi.org/10.3115/v1/W14-5302.
[5] R. Meyer, New wine in old wineskins?—Tagging Old Russian via annotation projection from modern translations, Russ. Linguist. 35 (2011) 267–281, http://dx.doi.org/10.1007/s11185-011-9075-x.
[6] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. Manning, S. Pyysalo, S. Schuster, F. Tyers, D. Zeman, Universal Dependencies v2: An evergrowing multilingual treebank collection, in: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association, Marseille, France, 2020, pp. 4027–4036, https://www.aclweb.org/anthology/2020.lrec-1.497.
[7] D.Q. Nguyen, K. Verspoor, An improved neural network model for joint POS tagging and dependency parsing, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text To Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 81–91, http://dx.doi.org/10.18653/v1/K18-2008.
[8] T. Tieleman, G. Hinton, Lecture 6.5-*rmsprop*: Divide the gradient by a running average of its recent magnitude, COURSERA Neural netw. mach. learn. 4 (2) (2012) 26–31.
[9] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Y. Bengio and Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, http://arxiv.org/abs/1412.6980.
[10] N. Pedrazzini, Exploiting cross-dialectal gold syntax for low-resource historical languages: Towards a generic parser for pre-modern Slavic, in: Proceedings of the Workshop on Computational Humanities Research (CHR 2020), CEUR Workshop Proceedings, Amsterdam, Netherlands, 2020, pp. 237–247, http://ceur-ws.org/Vol-2723/short48.pdf.
[11] D.T.T. Haug, M.L. Jøhndal, Creating a parallel treebank of the old Indo-European Bible translations, in: C. Sporleder, K. Ribarov, A. van den Bosch, M.P. Dobreva, M.J. Driscoll, C. Grover, P. Lendvai, A. Luedeling, M. Passarotti (Eds.), Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008), Marrakech, Morocco, 1 June 2008, ELRA, 2008, pp. 27–34, http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf#page=31.
[12] H.M. Eckhoff, K. Bech, G. Bouma, K. Eide, D.T.T. Haug, O.E. Haugen, M. Jøhndal, The PROIEL treebank family: A standard for early attestations of Indo-European languages, Lang. Resour. Eval. 52 (1) (2018) 29–65, http://dx.doi.org/10.1007/s10579-017-9388-5.
[13] S. Padó, M. Lapata, Dependency-based construction of semantic space models, Comput. Linguist. 33 (2007) 161–199, http://dx.doi.org/10.1162/coli.2007.33.2.161.
[14] Z. Ye, H. Zhao, Syntactic word embedding based on dependency syntax and polysemous analysis, Front. Inf. Technol. Electron. Eng. 19 (2018) 524–535, http://dx.doi.org/10.1631/FITEE.1601846.
[15] J. Boyd-Graber, D.M. Blei, Syntactic topic models, Comput. Linguist. 1 (1) (2006).
[16] S. Morsy, G. Karypis, Accounting for language changes over time in document similarity search, ACM Trans. Inf. Syst. 35 (1) (2016) http://dx.doi.org/10.1145/2934671.
[17] K. Berberich, S.J. Bedathur, M. Sozio, G. Weikum, Bridging the terminology gap in web archive search, in: 12th International Workshop on the Web and Databases, WebDB, 2009, http://webdb09.cse.buffalo.edu/papers/Paper20/webdb2009-final.pdf.
[18] H. Holzmann, G. Gossen, N. Tahmasebi, *Fokas*: Formerly known as – a search engine incorporating named entity evolution, in: Proceedings of COLING 2012: Demonstration Papers, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 215–222, https://www.aclweb.org/anthology/C12-3027.
[19] J.-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, E.L. Aiden, Quantitative analysis of culture using millions of digitized books, Science 331 (6014) (2011) 176–182, http://dx.doi.org/10.1126/science.1199644.
[20] L. Frermann, M. Lapata, A Bayesian model of diachronic meaning change, Trans. Assoc. Comput. Linguist. 4 (2016) 31–45, https://www.aclweb.org/anthology/Q16-1003.pdf.
[21] O. Popescu, C. Strapparava, SemEval 2015, Task 7: Diachronic text evaluation, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEVal 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 870–878, http://dx.doi.org/10.18653/v1/S15-2147.
[22] N.K. Tran, A. Ceroni, N. Kanhabua, C. Niederée, Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, in: WSDM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 339–348, http://dx.doi.org/10.1145/2684822.2685315.
[23] G. Chiron, A. Doucet, M. Coustaty, J.-P. Moreux, ICDAR2017 Competition on Post-OCR text correction, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 01, 2017, pp. 1423–1428, doi:10.1109/ICDAR.2019.00255.
[24] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, H. Margetts, Challenges and frontiers in abusive content detection, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 80–93, http://dx.doi.org/10.18653/v1/W19-3509.
[25] N. Tahmasebi, L. Borin, A. Jatowt, Survey of computational approaches to lexical semantic change, 2019, arXiv:1811.06278.
[26] X. Tang, A state-of-the-art of semantic change computation, Nat. Lang. Eng. 24 (5) (2018) 649–676, http://dx.doi.org/10.1017/S1351324918000220.
[27] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, *spaCy*: Industrial-strength Natural Language Processing in Python, Zenodo, 2020, http://dx.doi.org/10.5281/zenodo.1212303.