Exploiting Cross-Dialectal Gold Syntax for Low-Resource Historical Languages: Towards a Generic Parser for Pre-Modern Slavic

Nilo Pedrazzini[a]

[a] *University of Oxford, St Hugh's College, St Margaret's Rd, OX2 6LE, Oxford, United Kingdom*

Abstract

This paper explores the possibility of improving the performance of specialized parsers for pre-modern Slavic by training them on data from different related varieties. Because of their linguistic heterogeneity, pre-modern Slavic varieties are treated as low-resource historical languages, whereby cross-dialectal treebank data may be exploited to overcome data scarcity and attempt the training of a variety-agnostic parser. Previous experiments on early Slavic dependency parsing are discussed, particularly with regard to their ability to tackle different orthographic, regional and stylistic features. A generic pre-modern Slavic parser and two specialized parsers – one for East Slavic and one for South Slavic – are trained using jPTDP [8], a neural network model for joint part-of-speech (POS) tagging and dependency parsing which had shown promising results on a number of Universal Dependency (UD) treebanks, including Old Church Slavonic (OCS). With these experiments, a new state of the art is obtained for both OCS (83.79% unlabelled attachment score (UAS) and 78.43% labelled attachment score (LAS)) and Old East Slavic (OES) (85.7% UAS and 80.16% LAS).

Keywords

low-resource languages, dependency parsing, neural networks, early Slavic

1. Parsing data-poor historical languages: the case of Slavic

Low-resource languages represent a considerable challenge in Natural Language Processing (NLP), which is notoriously data-demanding. Data-poor languages and Big Data thus generally call for very different methodologies. Some languages may be 'low-resource' because they have only recently been recorded for the first time, with the ensuing difficulty of dealing with an ad hoc writing system or no standard orthography altogether, as is the case for many currently endangered languages [6]. Others may be widely spoken or relatively well-documented, but hardly have a sufficient amount of structured data (e.g. large, manually labelled corpora) to be target languages in downstream NLP tasks (e.g. [11]).

Low-resource historical languages present additional hurdles: they are closed sets and they necessarily lack native-speaker inputs. Unlike low-resource languages which can rely on data collection by means of fieldwork or on the expansion of manually annotated corpora by native speakers, the creation of structured data for historical languages is dependent on the digitisation of written sources that are virtually never only found in a contained geographic area. Even when digital editions are available, historical languages often lack a unified literary standard,

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands

nilo.pedrazzini@ling-phil.ox.ac.uk (N. Pedrazzini)
0000-0003-3757-2961 (N. Pedrazzini)

Table 1
Early Slavic dialect macro-areas and subvarieties with gold syntactic annotation in the TOROT Treebank

Dialect macro-areas	Varieties	Label	Tokens
South Slavic	Old Church Slavonic	OCS	139,055
	Serbian Church Slavonic	SCS	890
	Russian Church Slavonic	RCS	331
East Slavic	Old East Slavic	OES	142,138
	Middle Russian	MRus	95,066
	Old Novgorodian	ONov	2,245

which may result in linguistic and orthographic inconsistencies across and within individual texts. High orthographic variation, for instance, is obviously not ideal for the training of NLP models of a language which is already under-resourced to begin with.

Pre-modern Slavic varieties[1] are illustrative in this respect. Two dialect macro-areas, East and South Slavic, can be distinguished ever since the earliest Slavic sources (10-11th century) on the basis of various phonological and morphological features. However, the subvarieties belonging to each group have often very distinct characteristics. In an ideal world, the development of powerful tools for the processing of each dialect area would be carried out by using an equally large amount of data from all varieties. Table 1 shows which pre-modern Slavic varieties in the TOROT Treebank [4] [5] contain morphological and dependency annotation that could potentially be exploited for the development of NLP tools[2]. Not only is there a disproportion between the two dialect areas (with East Slavic being preponderant), but their subvarieties are far from being evenly distributed. Some major early Slavic varieties are not represented at all (e.g. Middle Bulgarian), which is also due to the fact that manual annotation can be slower or faster depending on the amount of secondary sources that may help speed up the process (e.g. translations and critical editions).

Computational techniques for the processing of early Slavic sources have been developing relatively quickly, including tools tackling the issue of obtaining digital primary sources (e.g. neural networks for handwritten text recognition [13]). The state of the art in automatic POS and morphological tagging has also reached success rates nearly as high as those of contemporary-language taggers [14]. By contrast, syntactic annotation is still performed almost exclusively manually. The result is that several texts in the corpus contain either no gold dependency annotation, or morphological tagging only. This arguably defies the very purpose of digital corpora of low-resource languages, which may be expected to contain detailed annotation throughout, precisely because they are necessarily limited in size. Besides, the implementation of annotation schemes pertaining to linguistic levels deeper than syntax (e.g. information and discourse structure) fully relies on having high-quality syntactic annotation

[1] 'Pre-modern Slavic' here assumes early Slavic varieties of the so-called Slavia Orthodoxa [12], that is, chiefly excluding all West Slavic languages (the Slavic subgroup which includes contemporary Czech, Slovak and Polish).

[2] A detailed breakdown of all the texts in corpus (including the labels with which they are referred to throughout the paper), with an indication on their language variety and number of tokens, can be found in the Appendix.

in the first place. Even more importantly, syntactically annotated corpora can be exploited for corpus-driven typological analyses, which can be crucial to advance linguistic theory. The disparity between low- and high-resource languages with regard to the availability of such resources thus risks to generate a bias towards patterns observed in the latter.

1.1. State of the art in pre-modern Slavic dependency parsing

Previous attempts at developing parsers for pre-modern Slavic have only been carried out on one of its dialect areas or on specific subvarieties:

- In [3], a parser for OES was trained using MaltParser [9] and was shown to be an efficient pre-annotation tool, yielding a decent annotation speed gain, but with a considerable difference between experienced and inexperienced annotators. However, as the authors note, its best scores (84.5% UAS and 73.4% LAS) were likely due to the simple genre and to the few long sentences of the test set. To the best of my knowledge, the experiment still represents the state of the art in the automatic parsing of OES.

- An off-the-shelf parser for OCS is instead available from UD [10][3]. The model, which reached relatively high scores (80.6% UAS and 73.4% LAS) was however only trained and tested on a single text (viz. MARIANUS). As a result, these scores do not reflect real-world performance[4] and the parser is hardly applicable to texts falling outside the set of orthographic and linguistic peculiarities of MARIANUS, which are only shared to some extent by the other texts classified as 'OCS' in Table 1.

- Finally, a neural network model has recently been trained on a number of UD treebanks, including OCS, using bidirectional long-short memory (BiLSTM) to jointly learn POS tagging and dependency parsing [8] (jPTDP). Its results for dependency parsing are similar to those of the off-the-shelf UD baseline OCS parser, but with a slight LAS improvement (+0.5%), thus representing the state of the art for OCS. Nevertheless, the same issue pointed out about the UD baseline parser applies: the scores given in [8] only refer to MARIANUS, which makes the model unusable beyond OCS texts that present orthographic and linguistic features very close to those of MARIANUS itself.

1.2. Aims of this paper

The goal of this paper is twofold:

- To investigate the extent to which the performance of specialized (i.e. variety-specific) parsers can be improved by expanding the training set with data from other varieties and dialect areas.

- To explore the possibility of attaining a 'generic' parser, a tool which is relatively dialect-agnostic and more flexible with respect to genres and historical stages.

A generic parser could especially speed up the annotation of pre-modern Slavic texts whose language and orthography are not straightforwardly classifiable in terms of provenance. Early

[3] http://ufal.mff.cuni.cz/udpipe/models

[4] In this context, 'real-world performance' refers to how well a model deals with texts that present different orthographic, regional and stylistic features

Slavic texts written in hybrid varieties are in fact rather the norm than the exception, which is due to intricate manuscript traditions, to the lack of a unified written standard, and to the complex relationship between vernacular(s) and literary language(s).

This experiment attempts to enhance the real-world performance of jPTDP [8], by training it on three different datasets: one containing only South Slavic data (OCS, RCS and SCS), one only East Slavic data (OES, MRus and ONov), and one both macro-varieties. The choice of retraining jPTDP rather than attempting to develop a novel neural network model is motivated, on the one hand, by the fact that jPTDP seems to perform particularly well with morphology-rich languages, which is the case for Slavic; on the other hand, we are interested in noting how the addition of heterogeneous training data affects its performance on OCS, which is the only pre-modern Slavic variety represented among the UD treebanks. Besides, jPTDP has not been tested on early East Slavic data, which allows us to compare the performance of a neural network model to that of MaltParser.

Section 2 outlines the pre-processing stage, including the criteria used to split the corpus into training, development and test sets. Section 3 lays out the training of jPTDP, including the choice of hyperparameters, and compares the results obtained for the three parsers during cross-validation. Section 4 is dedicated to the evaluation of the parsers by means of test sets which are meant to be indicative of real-world performance. Conclusions then follow with suggestions for future experiments.

2. Pre-processing

All the data used in this experiment comes from the latest TOROT Treebank release[5]. The corpus includes pre-modern Slavic text spanning from the oldest Slavic attestations (10th-11th century) to OES and MRus texts from the 11th-19th century [1]. It also includes a contemporary Russian subcorpus, which was however left out since we are only interested in the early stages of Slavic.

In order to reach representativeness and limit overfitting, 10% of each text was set aside as development data (40,375 tokens), 10% as test data (39,886 tokens) and 80% as training data (240,571 tokens). Texts with fewer than 400 tokens were exclusively employed for training[6]. By doing so, we obtained a relatively homogeneous distribution of genres and language varieties. Only for MARIANUS the predefined UD split into training, development and test set was adopted, to allow a comparison between our results and those of [8].

The training, development and test portions of each text are kept separate and merged only at need, which allows for faster experimentation with different combinations of texts while keeping the proportions consistent throughout[7].

TOROT releases come in two formats: the standard PROIEL XML format and the CoNLL-X format of UD. jPTDP requires the updated CoNLL-U format as input, whose main differences with CoNLL-X are the treatment of multiword tokens as integer ranges and the insertion of comments before each new sentence, besides the different order and outlook of their morphotags (e.g. NUMBs|GENDn|CASEn in CoNLL-X and Case=Nom|Gender=Neut|Number=Sing in CoNLL-U). The datasets were converted from PROIEL XML to CoNLL-U using the script

[5] https://github.com/torottreebank/treebank-releases/releases/tag/20200116

[6] This number (i.e. 400 tokens) was simply the minimum which allowed to split each text with a 80:10:10 proportion.

[7] All the datasets used in this experiment can be found at https://doi.org/10.6084/m9.figshare.12950093.v1. These include separate training, development and test files for each individual text.

Table 2
Highest scores obtained during cross-validation using the optimal hyperparameters for each dataset

Model	LSTM hidden states size	MLP hidden layer size	LAS	UAS
jDPTD-SSL	128	300	71.10	78.95
jDPTD-ESL	128	300	73.65	79.95
jDPTD-GEN	256	200	72.07	79.39

included in the Ruby utility `proiel-cli`, which can be used for the manipulation of PROIEL treebanks[8].

3. Training

In the first round of training, jPTDP was applied directly off-the-shelf with its default hyperparameters, in order to compare the scores in [8] with those resulting from our larger training set: 30 training epochs, 50-dimensional character embeddings, 100-dimensional word embeddings, 100-dimensional POS tag embeddings, 2 BiLSTM layers, 128-dimensional LSTM hidden states and 100 hidden nodes in each one-hidden-layer multi-layer perceptron (MLP). The hyperparameters were thus set by the authors of [8] on the basis of the optimal hyperparameters for the English WSJ Penn Treebank [7], which they established through a minimal grid search and applied to all UD treebanks without individual optimization. The only exception is the default size of LSTM hidden states, which they fixed at 128, even though the optimal value on the English WSJ Penn Treebank was found to be 256.

In the second round of training a grid search was performed to select the optimal size of LSTM hidden states in each layer from {128, 256} and the number of hidden nodes in MLPs from {100, 200, 300}. Due to limited computational resources, the other hyperparameters were set to default.

While the experiment in [8] suggests a better performance for jPDTP using 256-dimensional LSTM hidden states, our results during cross-validation indicate that this is not necessarily the case with pre-modern Slavic data. As Table 2 shows, only the generic model (jDPTD-GEN) benefits from a larger number of BiLSTM dimensions, whereas the specialized models, both the South Slavic (jDPTD-SSL) and the East Slavic one (jDPTD-ESL), perform better with a larger number of hidden nodes in MLPs (300), but 128 BiLSTM dimensions.

In Section 4 separate evaluations of the models developed with default and optimized hyperparameters will be provided for the sake of comparison. The evaluation phase will also show not only that real-world performance varies greatly depending on the text, but also that the scores emerged during cross-validation do not reflect the relative quality of the trained parsers. In all likelihood, this is primarily due to the fact that the development sets are virtually fully homogeneous, linguistically and stylistically, with the respective training sets, since they both comprise a percentage of nearly all texts written in the relevant Slavic variety.

[8] https://github.com/proiel/proiel-cli

Table 3
Test sets description

Label	Dialect macro-area	Varieties	Texts
SS	South Slavic	OCS, SCS	All south Slavic texts (>400 tokens)
CM	South Slavic	OCS	MARIANUS
CS	South Slavic	OCS	SUPR
VC	South Slavic	SCS	VIT-CONST
ES	East Slavic	OES, MRus, ONov	All east Slavic texts (>400 tokens)
PC	East Slavic	OES	LAV
SR	East Slavic	MRus	SERGRAD
AV	East Slavic	MRus	AVV
ON	East Slavic	ONov	BIRCHBARK

4. Evaluation

Each parser was tested on nine datasets which were chosen as representative of distinct early Slavic varieties and historical stages (Table 3). In particular:

- SS and ES, containing 10% of all East and South Slavic text respectively, are meant to show the performance of the parsers on the relevant dialect macro-areas as a whole.
- CM corresponds to the test set of both the UD baseline OCS parser and [8].
- CS is used to compare the performance of the parsers on OCS texts other than MARIANUS. As a miscellany, the syntax of SUPR is more varied than MARIANUS, which exclusively contains OCS translations of the Gospels. Moreover, though both very archaic (i.e. relatively close to reconstructed Proto-Slavic), they present different regional features (Bulgarian in SUPR, Macedonian in MARIANUS) and reflect different manuscript traditions (MARIANUS is a Glagolitic manuscript, SUPR a Cyrillic one).
- VC is used as the only late South Slavic manuscript (16th century) with clear Serbian features.
- PC is one of the most important OES manuscripts and the test sets used by [2].
- SR and AV represent two distinct varieties of MRus. The language of the former is in fact often classified as RCS, because of its hybrid Church Slavonic and Russian features. The latter is instead a 17th-century Russian text heavily influenced by the vernacular language.
- ON is representative of ONov, which is not only notoriously distinct from the Old Kievan and Moscovite varieties of early East Slavic (OES/MRus), but it also mostly consists of vernacular material – as opposed to the remaining east Slavic texts in the corpus, often heavily influenced by Church Slavonic (i.e. South Slavic).

The evaluation script which was used to compare gold and predicted tags can be found in the official UD repository[9].

[9] https://universaldependencies.org/conll18/evaluation.html

Table 4
Models evaluation: UAS-d[efault] and LAS-d[efault] are the scores obtained from the models trained with default hyperparameters, whereas UAS and LAS are those obtained from the optimized models, as defined in Table 2.

Test Set	Model	UAS	LAS	UAS-d	LAS-d
SS	jDPTD-SSL	76.99	69.51	77.08	69.54
	jDPTD-ESL	72.94	62.61	61.29	45.11
	jDPTD-GEN	**78.86**	**71.87**	78.11	70.72
CM	jDPTD-SSL	83.61	77.98	83.19	77.63
	jDPTD-ESL	83.60	77.83	66.03	50.98
	jDPTD-GEN	**83.79**	**78.42**	83.32	77.79
CS	jDPTD-SSL	68.54	58.76	69.21	59.30
	jDPTD-ESL	58.92	42.88	54.62	37.13
	jDPTD-GEN	**72.28**	**63.38**	71.22	61.53
VC	jDPTD-SSL	61.54	51.28	60.26	51.28
	jDPTD-ESL	66.67	48.72	60.26	43.59
	jDPTD-GEN	**69.23**	**56.41**	61.54	50.00
ES	jDPTD-SSL	62.83	47.55	63.18	47.51
	jDPTD-ESL	**81.02**	**74.93**	80.74	74.44
	jDPTD-GEN	80.86	74.23	80.67	74.18
PC	jDPTD-SSL	68.08	52.08	66.52	50.86
	jDPTD-ESL	**85.70**	**80.16**	85.59	79.25
	jDPTD-GEN	85.22	79.29	84.51	78.69
SR	jDPTD-SSL	58.63	41.42	58.84	41.16
	jDPTD-ESL	71.59	63.91	71.34	63.50
	jDPTD-GEN	73.24	64.71	**73.90**	**65.76**
AV	jDPTD-SSL	62.08	45.25	61.01	44.45
	jDPTD-ESL	80.91	74.96	79.44	73.89
	jDPTD-GEN	**81.75**	**75.80**	81.22	75.44
ON	jDPTD-SSL	58.82	41.18	57.75	41.18
	jDPTD-ESL	**74.33**	**58.82**	71.66	55.61
	jDPTD-GEN	72.19	58.29	68.98	53.48

As Table 4 shows, jDPTD-GEN performs best on all South Slavic test sets (SS, CM, CS, VC), as well as on the East Slavic AV and SR datasets. However, even when jDPTD-ESL performs better (viz. on ES, PC and ON) jDPTD-GEN does not lag far behind. This clearly indicates that cross-dialectal training data may improve the performance of a parser, even if it is meant to be used to annotate only text of a particular variety.

Unsurprisingly, there are also obvious indications that the level of representativeness of a variety among the training data has important consequences on the performance of the parsers. The scores obtained on VC and ON are particularly low, with LAS < 60.00. This is likely due to the fact that SCS and ONov linguistic features are greatly underrepresented in the corpus. Expanding the training data with these varieties is therefore very likely to improve the performance of the models.

Several errors are due to orthographic idiosyncrasies of the individual manuscripts (or the edition thereof), whereby an unusual spelling may render the syntactic relation of a word ambiguous. In (1), for instance, the word ѿтинȣд 'not at all' is spelt differently from its most usual forms, отиноуდь or ѡтинудь. It is likely that the parser expected a singular subject in its position, given the following main verb воздръжаше '(he) abstained'. The lack of final -ь in ѿтинȣд does in fact make the word appear morphologically like a singular masculine noun.

(1) и ѿ пїѧньс̆тва ѿтинȣд воздръжаше сѧ
and.CC from.CASE drinking.OBL not-at-all.ADVMOD abstain.ROOT himself.AUX
and.CC from.CASE drinking.OBL not-at-all.NSUBJ abstain.ROOT himself.AUX

(Gold)
(Predicted)
'And by no means did he abstain from drinking' (SERGRAD, 17r)

It is particularly interesting to note that the scores obtained on CS, the only other OCS text in the corpus, are not as high as those reached with CM, which corresponds to the test set in [8]. This is indicative of the fact that the previous state of the art in parsing OCS does not reflect real-world performance. In our case, the relatively low scores obtained on CS appear to be mostly due to its more complex syntactic structures compared to CM. In (2), for instance, the indirect object is repeated twice, once after the subject and once just before the main verb, which is plausibly the main cause of the poor performance of the parser on the rest of the sentence:

(2) й ѥ҆лико тебѣ любо й
and.CC whoever.NSUBJ to-you.IOBJ beloved.NSUBJ and.CC
and.CC whoever.ADVMOD to-you.ADVMOD beloved.XCOMP and.ADVMOD
драго тебѣ бѫде
dear.CONJ to-you.IOBJ will-be.ROOT (Gold)
dear.XCOMP to-you.OBL will-be.ROOT (Predicted)
'And whoever will be beloved and dear to you' (SUPR, 24v)

As Tables 5 and 6 show, with our models we obtained a new state of the art in both OCS and OES dependency parsing. It is worth noting that while jDPTD-ESL performed slightly better on PC, jDPTD-GEN is also past the state of the art for OES. This is a particularly promising result: as already discussed, because of the lack of standardization in pre-modern Slavic, the development of a high-quality generic parser should arguably be prioritized over that of multiple specialized models. While this could mean a slightly lower performance than specialized parsers when it comes to well-represented varieties (e.g. OES), the long-term benefit of a dialect-agnostic tool are likely to be more substantial. A decent-quality generic parser could in fact be employed to speed up the annotation of underrepresented varieties, which would ultimately result in the expansion of deeply annotated treebanks.

5. Conclusions and future experiments

This paper explored the possibility of exploiting syntactically annotated treebanks of related but distinct pre-modern Slavic varieties in order to train a generic, variety-agnostic parser.

Table 5
OCS: comparison with previous experiments on MARIANUS

Model	UAS	LAS
UD baseline model	80.6	73.4
jPTDP [8]	80.59	73.93
jDPTD-GEN (default hyperparameters)	83.32	77.79
jDPTD-GEN (optimized)	**83.79**	**78.42**

Table 6
OES: comparison with previous experiments on LAV

Model	UAS	LAS
Maltparser [2]	84.5	77.9
jDPTD-ESL (default hyperparameters)	85.59	79.25
jDPTD-ESL (optimized)	**85.7**	**80.16**

The results suggest that the performance of a specialized model can in fact be considerably improved by expanding the training data with different pre-modern Slavic varieties. This has particularly emerged from the scores obtained on OCS (South Slavic) by the generic parser, which was trained on both South and East Slavic data. With our experiment a new state of the art has been obtained for both OCS (UAS 83.79% and LAS 78.43%) and OES (UAS 85.7% and LAS 80.16%). Future studies may wish to attempt larger-scale experimentation with cross-lingual transfer learning across different related historical languages. The automatic processing of OCS is especially very likely to benefit from direct transfer or annotation projection, as well as from cross-lingual word representations, from Ancient and New Testament Greek, given the comparatively very similar linguistic systems of Slavic and Greek.

Acknowledgments

This work was supported by the Economic and Social Research Council [grant number ES / P000649 / 1]. I am grateful to Marius Jøndal for his help with Ruby while setting up the `proiel-cli` utility.

References

[1] A. Berdičevskis and H. M. Eckhoff. “A Diachronic Treebank of Russian Spanning More Than a Thousand Years.” In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020).* Marseille, France: European Language Resources Association, May 2020, pp. 5251–5256. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.646.

[2] H. Eckhoff and A. Berdicevskis. *Replication Data for: Automatic parsing as an efficient pre-annotation tool for historical texts.* Version V2. 2016. DOI: 10.18710/FERT42.

[3] H. M. Eckhoff and A. Berdičevskis. “Automatic parsing as an efficient pre-annotation tool for historical texts.” In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 62–70. URL: https://www.aclweb.org/anthology/W16-4009.

[4] H. M. Eckhoff and A. Berdičevskis. “Linguistics vs. Digital Editions: The Tromsø Old Russian and OCS Treebank.” In: *Scripta & E-Scripta* 14-15 (2015), pp. 9–25.

[5] H. M. Eckhoff et al. “The PROIEL treebank family: A standard for early attestations of Indo-European languages.” en. In: *Language Resources and Evaluation* 52.1 (2018), pp. 29–65.

[6] J. Essegbey. “‘Is this my language?’: Developing a writing system for an endangered-language community.” en. In: *Language Documentation and Endangerment in Africa*. Ed. by J. Essegbey, B. Henderson, and F. McLaughlin. Amsterdam: John Benjamins, 2015, pp. 153–176. DOI: 10.1075/clu.17.06ess.

[7] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. “Building a Large Annotated Corpus of English: The Penn Treebank.” In: *Computational Linguistics* 19.2 (1993), pp. 313–330. URL: https://www.aclweb.org/anthology/J93-2004.

[8] D. Q. Nguyen and 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*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 81–91. DOI: 10.18653/v1/K18-2008.

[9] J. Nivre, J. Hall, and J. Nilsson. “Maltparser: A data-driven parser-generator for dependency parsing.” In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. 2006, pp. 2216–2219. URL: http://lrec-conf.org/proceedings/lrec2006/pdf/162_pdf.pdf.

[10] J. Nivre et al. “Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection.” In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France: European Language Resources Association, 2020, pp. 4027–4036. ISBN: 979-10-95546-34-4.

[11] A. K. Ojha and D. Zeman. “Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri.” In: *Proceedings of the LREC 2020 WILDRE5 – 5th Workshop on Indian Language Data: Resources and Evaluation*. Paris, France: European Language Resources Association, 2020, pp. 33–38. ISBN: 979-10-95546-67-2.

[12] R. Picchio. *Letteratura della Slavia ortodossa: IX-XVIII sec*. Storia e Civiltà - Edizioni Dedalo. Dedalo, 1991. ISBN: 9788822005304. URL: https://books.google.co.uk/books?id=xVQjAnu-u0UC.

[13] A. Rabus. “Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus.” en. In: *Scripta & E-Scripta* 19 (2019).

[14] Y. Scherrer, A. Rabus, and S. Mocken. “New Developments in Tagging Pre-modern Orthodox Slavic Texts.” nl. In: *Scripta & E-Scripta* 18 (2018), pp. 9–33.

A. Dataset breakdown

Table 7
Dataset breakdown, with an indication of the language variety represented by each manuscript. The text labels reproduce the codes used by the official TOROT releases, to facilitate text retrieval should one wish to check the results of this paper against the original datasets.

Variety	Text	Label	Tokens
OCS	Codex Marianus	MARIANUS	58,269
	Codex Suprasliensis	SUPR	79,070
	Codex Zographensis	ZOGR	1,098
	Kiev Missal	KIEV-MIS	370
	Psalterium Sinaiticum	PSAL-SIN	248
SCS	Vita Constantini	VIT-CONST	890
RCS	Vita Methodii	VIT-METH	331
OES	Primary Chronicle (Codex Laurentianus)	LAV	56,725
	Suzdal Chronicle (Codex Laurentianus)	SUZ-LAV	23,760
	Primary Chronicle (Codex Hypathianus)	PVL-HYP	3,610
	First Novgorod Chronicle (Synodal)	NOV-SIN	17,838
	Kiev Chronicle (Codex Hypathianus)	KIEV-HYP	544
	Colophon (Mstislav's Gospel)	MSTISLAV-COL	259
	Colophon (Ostromir Codex)	OSTROMIR-COL	199
	Missive (Archbishop of Riga)	RIG-SMOL1281	171
	Mstislav's letter	MST	158
	Novgorod's treaty with Jaroslav	NOVGOROD-JAROSLAV	423
	Russkaja pravda	RUSPRAV	4,174
	Statute of Prince Vladimir	UST-VLAD	495
	Treaty (Smolensk-Riga-Gotland)	RIGA-GOTH	1,421
	The Tale of Igor's Campaign	SPI	2,850
	Russkaja pravda	RUSPRAV	4,174
	Uspenskij Sbornik (excerpts)	USP-SBOR	25,189
	Varlaam Xutynskij's Grant Charter	VARLAAM	148
MRus	Afanasij Nikitin's *Journey*	AFNIK	6,842
	Charter of Prince Jurij Svjatoslavich	SMOL-POL-LIT	344
	Correspondence of Peter the Great	PETER	100
	Domostroj	DOMO	23,459
	Life of Sergij of Radonezh	SERGRAD	20,361
	History of the schism (materials)	SCHISM	1,835
	Missive (Ivan of Pskov)	PSKOV-IVAN	339
	Testament (Ivan Jur'evič Graznoj)	DUX-GRAZ	421
	Life of Avvakum	AVV	22,835
	Tale of Dracula	DRAC	2,487
	The tale of Luka Koločskij	LUK-KOLOC	906
	The taking of Pskov	PSKOV	2,326
	The tale of the fall of Constantinople	CONST	9,258
	Vesti-Kuranty	VEST-KUR	1,154
	Zadonščina	ZADON	2,399
ONov	Birchbark letters	BIRCHBARK	1,965
	Novgorod service book marginalia	NOV-MAR	93
	Novgorodians' losses	NOV-LIST	187