sciendo
DOI 10.2478/jazcas-2023-0040

THE EFFECT OF (HISTORICAL) LANGUAGE VARIATION ON THE EAST SLAVIC LECTS LEMMATISERS PERFORMANCE

ILIA AFANASEV – OLGA LYASHEVSKAYA
– STEFAN REBRIKOV – YANA SHISHKINA
– IGOR TROFIMOV – NATALIA VLASOVA
Independent researchers

AFANASEV, Ilia – LYASHEVSKAYA, Olga – REBRIKOV, Stefan – SHISHKINA, Yana – TROFIMOV, Igor – VLASOVA, Natalia: The Effect of (Historical) Language Variation on the East Slavic Lects Lemmatisers Performance. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 225 – 233.

Abstract: The need to develop tools for historical and regional variations is becoming more urgent in natural language processing. In this paper, we present two candidate systems for lemmatising historical East Slavic lects (Late Old East Slavic and Middle Russian), as well as modern regional East Slavic lects (Belogornoje and Megra): BERT-based end-to-end pipeline with language-specific heuristics and sequence-to-sequence BART-based encoder-decoder. To evaluate their predictions, we use accuracy score and string similarity measures, such as Levenshtein distance. The BERT-based model is more suitable for the regional data, achieving 85% accuracy score, and only 74% on the historical data. BART-based model climbs up to 92.6% accuracy score on the historical data, yet gets only 80% on the regional data. We provide an error analysis and discuss ways to enhance models, such as dictionary lookup and spellchecker.

Keywords: East Slavic, language variation, lemmatisation, dialectology, historical linguistics, historical NLP

1 INTRODUCTION

Lemmatisation is an NLP task that includes predicting the dictionary form of a token, a lemma. Lemmatisation is crucial for the linguistic downstream tasks, such as dictionary compilation, or grammar description (Straka – Straková 2017; Bergmanis – Goldwater 2018; Kanerva et al. 2021), especially for historical and regional lects (Berdičevskis et al. 2016; de Graaf et al. 2022). However, the impact of language variation on the performance of lemmatisers is still understudied.

In this paper, we examine how two types of language variation, historical and regional, challenge two types of lemmatisers: a transformer-based end-to-end

pipeline system with heuristics that enhance its performance on a standard modern language (Anastasyev 2020), and a sequence-to-sequence transformer-based lemmatiser (Lewis et al. 2020).

We study two types of variation that may affect the given lemmatisers' performance. Historical material consists of Late Old East Slavic and Middle Russian texts. The northern and southern Russian territorial lects material represents the regional data. We hypothesise the following:

H1: Prioritising rare words (and, subsequently, non-productive word-formation models) enhances the performance of the language-specific pipeline model lemmatiser module over the regional data.

H2: Variation within the training historical data helps the sequence-to-sequence transformer-based lemmatiser to demonstrate better results on the historical evaluation subset. In addition, we are going to examine whether preliminary morphological tagging enhances the results of the sequence-to-sequence lemmatiser.

Section 2 discusses the previous research on the topics of East Slavic NLP and lemmatiser development current trends. In Section 3 we describe the training datasets and the test data. Section 4 presents the models used in the study. In Section 5 we report the conducted experiments and discuss their results. Section 6 wraps up the current research and sketches its further direction.

2 RELATED WORK

The most widely used lemmatisers are multilingual, with the underlying inference engines being language-agnostic (Straka – Straková 2017; Bergmanis – Goldwater 2018; Kanerva et al. 2021). These tools generally utilize sequence-to-sequence architecture (Sutskever et al. 2014; Cho et al. 2014) implemented via encoder-decoder transformers such as BART (Lewis et al. 2020).

The current trends stimulate the researchers of particular languages and language groups to develop the lemmatisers for a particular lect or a set of lects (Anastasyev 2020; Fernández 2020). The scholars employ external resources such as dictionaries (Milintsevich – Sirts 2021), thus significantly improving the models' efficiency for extremely low-resourced lects (de Graaf et al. 2022).

East Slavic NLP for a long period concentrated around standard Russian (Anastasyev 2020) and its historical varieties (Berdičevskis et al. 2016; Lyashevskaya – Penkova 2021; Pedrazzini – Eckhoff 2021). However, the particular texts that are the focus of this study were digitalised fairly recently and were not tagged (Novokhatko 2012; Kusmina – Filippova 2013).

The interest in standard Ukrainian (Omelianchuk et al. 2020; Omelianchuk et al. 2021) and standard Belarusian (Shishkina – Lyashevskaya 2021) sparked in recent years. However, there is hardly any information on attempts to develop NLP tools for smaller modern East Slavic lects, despite, for instance, Saratov dialectological school's

active study of the Belogornoje and Megra lects and the development of a corpus and dialectal dictionaries (Kruchkova – Goldin 2011; Kruchkova – Goldin 2015).

3 DATA

The first training dataset consists of different standard Modern Russian texts balanced by genres (news, poetry, fiction, social media, the latter taken from the Taiga corpus (Shavrina – Shapovalova 2017)), orthography (premodern/modern), and periods (the 1700s, 1800s, and 1900s–2020s): overall, ca. 2,000,000 tokens.

The second training dataset consists of the Late Old East Slavic and Middle Russian legal texts from 1400 to 1700 taken from different collections (Russian History Library 1875; Rozysknyje dela o Fedore Shaklovitom i jego soobshchnikakh 1893; Likhachov 1954; Cherepnin 1961; Ankhimiuk 2000). The overall size of this collection is ca. 923,000 tokens.

The regional test data consist of two groups of East Slavic lects, northern Megra (Vologda Region, Russia), and southern Belogornoje (Saratov Region, Russia). The texts spelt close to the actual pronunciation of the native speakers are taken from the Saratov dialectological corpus (Kruchkova – Goldin 2011). Together they form a dataset of ca. 8,000 tokens.

The historical test data consist of a single set of documents, Besobrasow's archive (Novokhatko 2012; Kusmina – Filippova 2013). These are legal texts of the latter half of the 1600s. The overall size of these texts achieves ca. 437,000 tokens.

4 METHOD

In this research, we compare two approaches to the lemmatisation of historical and regional East Slavic varieties.

The first approach is a robust end-to-end pipeline model, which performs morphological tagging, lemmatisation, and dependency parsing. We use qbic (Anastasyev 2020), a state-of-the-art ensemble for Russian that makes use of a pre-trained RuBERT model. We modify the ensemble with a RuBERT-large model; lemmatisation making use of morphological classification; dictionary lookup (a mapping of word with its part-of-speech to its lemma); correcting symbol sequences forbidden in Russian; and orthographical normalisation. We refer to this model as Rubic.

The second approach focuses on sequence-to-sequence lemmatisation that can use morphological information from data. We use only the BART-large pre-trained model without additional enhancements.[1]

[1] Source code is available at https://github.com/The-One-Who-Speaks-and-Depicts/transformer-lemmatiser; models are available at https://huggingface.co/djulian13/bart-large-Modern-Russian-lemmatisation and https://huggingface.co/djulian13/bart-large-Middle-Russian-lemmatisation.

The training generates four models: Rubic for modern data, BART-large for modern data, Rubic for historical data, BART-large for historical data. We use some heuristics for augmenting the modern training data: capitalisation, quotation marks-to-guillemets conversion, and jo-fication (this heuristic transforms e into ë in tokens when they are interchangeable by the standard Russian rules). These additional heuristics allow to add contexts for words from rare grammatical paradigms.

We test the models on historical and regional data. Initially, the conditions for Rubic and BART-large test on regional data are not equal. BART-large lemmatiser is trained with the use of morphological information, which the regional data lack. To solve this issue, we train the morphological tagger (Scherrer 2021) on standard Russian data. It achieves the 87% F1-score, and we use it to produce a silver (more or less reliable, but not reliably checked by humans) morphological tagging for the regional data.

For an initial evaluation, we use an accuracy score. We also implement Levenshtein (Levenshtein 1966), Damerau-Levenshtein (Damerau 1964), and Jaro-Winkler (Jaro 1989; Winkler 1990) distances, a method that allows us to get fine details (Lyashevskaya – Afanasev, in print), to study Rubic and BART-large performance on the regional data.

The analysis aims at understanding the key errors and the explanation of models' performance, both in comparison and on their own. The following discussion of the results is required to get an outline for future work on both approaches.

5 RESULTS AND ANALYSIS

The four models achieve an accuracy score from 85 to 99% (depending on the exact subset) on both modern and historical data. Two series of experiments are conducted. The first series of experiments include testing the Rubic and BART-large models, fine-tuned on the historical data, on Besobrasow's archive. The second series includes testing the Rubic and BART-large, fine-tuned on the modern data, on regional East Slavic lects material.

5.1 Experiment 1

Tab. 1 presents the results of the experiments on the Besobrasow's archive.

Dataset	Besobrasow's archive
Rubic (token normalisation)	73.8
BART-large	85.0
BART-large (token normalisation)	**92.6**

Tab. 1. The accuracy score, %, of the historical East Slavic datasets lemmatisation by Rubic and BART-large

Both models are sensitive to the orthographical variation: BART-large gets the 0.07 boost when the tokens are normalised. The sequence-to-sequence architecture performs better: the Rubic augmentations enhance its performance on standard Russian and not the historical material (Lyashevskaya et al. 2023).

BART-large performance is not perfect. The historical data are significantly more heterogeneous than the modern data, and it is harder for the model to predict non-standard inflexion because the chance to meet it in the training data is smaller. For comparison, an element -мѣрити is often replaced by a more frequent -мерети, in verbs like смѣрити 'measure', cf. умерети 'die'. Nouns and adjectives are also affected, cf. выклатка instead of выкладка 'facing' influenced by оплатка 'wafer', and друговъ instead of другой 'other', influenced by каковъ 'which'. There are also non-standard transformations, cf. еиц >> яйцо 'egg', for which BART-large correctly predicts the lemmatisation model, yet does not predict the in-root phonetic alternation producing the forms like ейцо. Sometimes, the specifics of the annotation schema cause errors. The earlier data use lemma и for a masculine 3rd person pronoun, and the later data uses the lemma онъ, which confuses the model, unaware of the East Slavic lects evolution.

5.2 Experiment 2

Next, we test the models' abilities to operate within the conditions of synchronic variation. The dataset presents an additional challenge for a sequence-to-sequence model. The morphological tags of the test data do not correspond to the training tag set, so we use silver tagging made with Scherrer (2021). We run the BART-large model separately on non-tagged and tagged data.

Tab. 2 and 3 demonstrate the results for the Megra and Belogornoje datasets.

Metrics	A	L	L: N	D-L	D-L: N	J-W	J-W: N
Rubic	**86.9**	**0.31**	**0.31**	**0.31**	**0.31**	**~98.0**	**~98.0**
BART-large	49.66	1.08	0.94	1.08	0.94	89.4	94.7
BART-large (pre-tagged data)	82.67	0.37	0.37	0.37	0.37	95.7	95.7

Tab. 2. The results of the Megra dataset lemmatisation by Rubic and BART-large: accuracy score (A), %; averaged string similarity measures (Levenshtein distance (L), Levenshtein distance, normalised (L:N), Damerau-Levenshtein distance (D-L), Damerau-Levenshtein distance, normalised (D-L: N); averaged Jaro-Winkler distance (J-W) and Jaro-Winkler distance, normalised (J-W: N)), %.

Metrics	A	L	L: N	D-L	D-L: N	J-W	J-W: N
Rubic	**87.8**	**0.25**	**0.25**	**0.25**	**0.25**	**98.2**	**98.2**
BART-large	52.97	1.02	0.85	1.02	0.85	91.3	96.1
BART-large (pre-tagged data)	84.89	0.29	0.29	0.29	0.29	97.9	97.9

Tab. 3. The results of the Belogornoje dataset lemmatisation by Rubic and BART-large: accuracy score (A), %; averaged string similarity measures (L, L: N, D-L, D-L: N); averaged Jaro-Winkler distances (J-W, J-W: N), %.

In this experiment BART-large performs worse than Rubic: the heuristics of the latter enable it to overcome some difficulties of the regional data lemmatisation, like ещё/еще contrast. Note that BART-large demonstrates dependency on preliminary morphological tagging: the model run on raw data falls behind the model run on tagged data by 30%.

Generally, the high results of string similarity measures show that each model captures the general principle of lemmatisation: the output, lemma, is usually similar to the token that is provided as a part of the input. The equal results of Levenshtein and Damerau-Levenshtein distances support the claim: they mean that there are no character transpositions in target/output pairs. The equal normalised and non-normalised string similarity measures for Rubic and BART-large, run on tagged data, demonstrate that the models generate sequences similar to the target and do not prioritise a particular inflexion type over the others. BART-large, run on raw data, on the other hand, often produces erroneous output.

There are some general issues that may explain why the models struggle with the regional data. Lemmata in gold data reflect a particular lect norm, not a standard norm. Thus, correctly – by standard means – predicted еще 'yet' becomes incorrect, when compared to ещё. The relatively high string similarity measures scores support the hypothesis of the relatively high rate of this kind of errors. Linguistic differences between modern standard Russian and regional East Slavic lects also negatively affect the models' performance. Models are not aware of some pronoun forms, and lemmatise, for instance, мни '1.SG.DAT' as мни and not я '1.SG.NOM'. Non-standard spelling, combined with non-productive paradigms, leads to errors: models predict робят instead of робёнок 'child-SG.NOM'. The gold data also contain multi-word lemmata which training data lacks, which also results in errata, cf. gold вот и and predicted и 'and'. These issues are well-known for the smaller lects lemmatisation: there is no consensus on what should be considered a lemma in such cases (de Graaf et al. 2022).

6 CONCLUSION

In this paper, we studied how the synchronic and diachronic variation in East Slavic languages impacts the lemmatisers' performance. We compared the lemmatiser module of a robust end-to-end BERT-based pipeline (Rubic) and the independent sequence-to-sequence BART-based model (BART-large). Different kinds of variation negatively affect the performance of both models, lowering their average accuracy score to 80–90% from more than 90% on the validation data. BART-large overcomes historical variation more easily, especially when provided with gold morphological tagging, which is supported by string similarity measures evidence. Rubic does not depend on morphological data and better overcomes the synchronic variation of the territorial East Slavic lects closely related to standard

Russian. Both models are confused by orthographical variation, non-standard paradigms, and difference of lemma choice in golden data.

Both models need enhancements. For BART-large, we could suggest implementing heuristics, in a similar fashion to Rubic's. Built-in lect identification may be of some help, normalising tokens before their lemmatisation. As symbol-by-symbol generation causes a lot of BART-large errors, this model may benefit from spellchecker or a dictionary lookup during the postprocessing. The biggest issue is the BART-large model being a stand-alone lemmatiser, it needs the morphological tagging complement to be efficient, in contrast with Rubic. The development of tagging tool should be the primary focus of future research. The same heuristics may be useful for Rubic, a model that is even more suitable for this kind of modification, as it already employs some heuristics. It may also benefit from other types of pre-trained language models, pushed up in its architecture.

The datasets augmentation approaches include incorporating irregular lemmatisation models, as well as the balancing of wordforms by the frequency of word changing paradigm. The modern dataset will benefit from the addition of a regional component.

References

Anastasyev, D. (2020). Exploring pretrained models for joint morphosyntactic parsing of Russian. In Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue", 19, pages 1–12, Moscow, Russia.

Ankhimiuk, U. V. (2000). Soligalicheskije akty iz "Arkhiva Volynskikh". In A. V. Antonov (ed.): Russian Diplomatary. Moscow: Archeographical center, pages 25–42.

Berdičevskis, A., Eckhoff, H., and Gavrilova, T. (2016). The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian. In Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog», pages 99–111, Moscow, Russia. RSSU.

Bergmanis, T., and Goldwater, S. (2018). Context sensitive neural lemmatization with Lematus. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Cherepnin, L. V. (1961). Akty feodal'nogo zemlievladenija i khozyajstwa XIV – XVI vekov (in 3 volumes). Moscow: USSR Academy of Sciences.

Cho, K., Merriënboer van, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3), pages 171–176.

Fernández, L. G. (2020). A contribution to Old English lexicography. NOWELE / North-Western European Language Evolution, 73(2), pages 236–251.

Graaf de, E., Stopponi, S., Bos, J. K., Peels-Matthey, S., and Nissim, M. (2022). AGILe: The first lemmatizer for Ancient Greek inscriptions. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5334–5344, Marseille, France. European Language Resources Association.

Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. Journal of the American Statistical Association, 84, pages 414–420.

Kanerva, J., and Ginter, F., and Salakoski, T. (2021). Universal lemmatizer: A sequence--to-sequence model for lemmatizing universal dependencies treebanks. Natural Language Engineering, 27(5), pages 545–574.

Kruchkova, O., and Goldin, V. (2011). Corpus of Russian dialect speech: concept and parameters of evaluation. In Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2011", pages 359–367, Moscow, Russia.

Kruchkova, O., and Goldin, V. (2015). The parameters of text processing for the Russian dialect corpus. In Proceedings of the international conference "Corpus linguistics — 2015", pages 307–314, Saint Petersburg, Russia.

Kuzmina, O. V., and Filippova, I. S. (2012). Arkhiv stol'nika Andreja Il'jicha Besobrasowa, vol. II. Moscow: Russian History Insitute of Russian Academy of Sciences, 877 p.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), pages 707–710.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Likhachov, D. S. (1954). Puteshestvija russkich poslov XVI – XVII vv. Statejnyje spiski. Moscow, Leningrad: USSR Academy of Sciences, 490 p.

Lyashevskaya, O., Afanasev, I., Rebrikov, S., Shishkina, Y., Suleymanova, E., Trofinov, I., and Vlasova, N. Disambiguation in context in the Russian National Corpus: 20 years later. In Proceedings of International Conference "Dialogue 2023", pages 1–12, Online.

Lyashevskaya, O., and Afanasev, I. (in print). String similarity measures for evaluating the lemmatisation in Old Church Slavonic. In Proceedings of International Conference on Historical Lexicography and Lexicology. La Rioja, Spain. Universidad de La Rioja.

Lyashevskaya, O., and Penkova, Y. (2021). Revised entries in the multi-volume edition and TEI encoding: a case of the historical dictionary of Russian. In Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, vol. II, pages 655–662. Komotini, Greece. Democritus University of Thrace.

Milintsevich, K., and Sirts, K. (2021). Enhancing sequence-to-sequence neural lemmatization with external resources. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3112–3122, Online. Association for Computational Linguistics.

Novokhatko, O. V. (2012). Arkhiv stol'nika Andreja Il'jicha Besobrasowa, vol. I. Moscow: Russian History Insitute of Rassian Academy of Sciences, 903 p.

Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanskyi, O. (2020). GECToR – Grammatical Error Correction: Tag, Not Rewrite. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 163–170, Seattle, WA, USA. Online. Association for Computational Linguistics.

Omelianchuk, K., Raheja, V., and Skurzhanskyi, O. (2021). Text Simplification by Tagging. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 11–25, Online. Association for Computational Linguistics.

Pedrazzini, N., and Eckhoff, H. M. (2021). OldSlavNet: A scalable Early Slavic dependency parser trained on modern language data. Software Impacts, 8, pages 1–4.

Rozysknyje dela o Fedore Shaklovitom i jego soobshchnikakh, in 4 volumes (1893). Saint Petersburg: Arkheological Commission.

Russian History Library, volume II (1875). Saint Petersburg: Arkheological Comission, 351 p.

Scherrer, Y. (2021). Adaptation of morphosyntactic taggers: Cross-lectal and multilectal approaches. In M. Zampieri – P. Nakov (eds).: Similar languages, varieties, and dialects: A computational perspective. Studies in Natural Language Processing, Cambridge University Press, pages 138–166.

Shavrina, T., and Shapovalova, O. (2017). To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. In Proceedings of the International Conference "CORPORA 2017", Saint-Petersbourg, Russia.

Shishkina, Y., and Lyashevskaya, O. (2021). Sculpting enhanced dependencies for Belarusian. In Revised Selected Papers of Analysis of Images, Social Networks and Texts: 10th International Conference (AIST 2021), pages 137–147. Tbilisi, Georgia.

Straka, M., and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014) Sequence to sequence learning with neural networks In Z. Ghahramani – M. Welling – C. Cortes – N. Lawrence – K. Q. Weinberger (eds.): Advances in neural information processing systems, vol. 27. Proceedings of NIPS 2014, pages 3104–3112, Montreal, Curran.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In Proceedings of the Section on Survey Research Methods, pages 354–359, Alexandria, VA. American Statistical Association.

Zaharova, K. F., and Orlova, V. G. (2004). Dialektnoe chlenenie russkogo yazyka. Moscow: URSS, 176 p.