Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

THE SMALLER THE BETTER? HETEROGENEITY OF CORPUS, TRAINING SIZE, AND MORPHOLOGICAL TAGGING

Lyashevskaya O. N. (olesar@yandex.ru),
Ostyakova L. N. (lostaaa15@gmail.com),
Salnikov E. A. (egorsalnikov1@gmail.com),
Semenova O. A. (olga.sem10@gmail.com)
HSE University, Moscow, Russia

Orthographic and morphological heterogeneity of historical texts in pre-modern Slavic causes many difficulties in pos- and morphological tagging. Existing approaches to these tasks show state-of-the-art results without normalization, but they are still very sensitive to the properties of training data such as genre and origin. In this paper, we investigate to what extent the heterogeneity and size of the training corpus influence the quality of pos tagging and morphological analysis. We observe that UDpipe trained on different parts of the Middle Russian corpus demonstrates a boost in accuracy when using less training data. We resolve this paradox by analyzing the distribution of pos-tags and short words across subcorpora.

Keywords: part of speach tagging, full morphological tagging, historical data, data size variation, data homogeneity

DOI: 10.28995/2075-7182-2020-19-1091-1108

МАЛ ДА УДАЛ? КАК ГЕТЕРОГЕННОСТЬ КОРПУСА И РАЗМЕР ТРЕНИРОВОЧНЫХ ДАННЫХ ВЛИЯЮТ НА АВТОМАТИЧЕСКУЮ МОРФОЛОГИЧЕСКУЮ РАЗМЕТКУ

Ляшевская О. Н. (olesar@yandex.ru),
Остякова Л. Н. (lostaaa15@gmail.com),
Сальников Е. А. (egorsalnikov1@gmail.com),
Семенова О. А. (olga.sem10@gmail.com)
НИУ ВШЭ, Москва, Россия

Гетерогенность орфографии и грамматического строя текстов старорусской письменности создают существенные трудности для автоматической частиречной и морфологической разметки. Существующие подходы демонстрируют хорошие результаты, не прибегая к помощи нормализации, однако все они, тем не менее, чувствительны к любым изменениям пропорций элементов тренировочного датасета и жанровой неоднородности. В данной работе мы проанализировали влияние этих факторов на качество автоматической морфологической разметки. Наше исследование показало, что качество морфологической разметки моделей UDpipe повышается по мере снижения объёма тренировочных данных. Именно поэтому нами была предпринята попытка проанализировать дистрибуцию частей речи и слов, состоящих из малого количества символов (2–3), в тренировочных выборках.

Ключевые слова: частеречная разметка, морфологическая разметка, обработка исторических текстов

1. Introduction

Middle Russian Corpus is a part of the Russian National Corpus that mostly represent historical texts created in the 15–17th centuries. Processing such historical corpora is a challenging task due to the absence of standard spelling and changes in grammatical structure over the period [Arkhangelsky et al. 2014]; [Gavrilova et al. 2017]. Middle Russian is a highly-inflected language, which results in a large tagset. Moreover, the language variation of Middle Russian grammar and lexicon depending on schools, manuscripts and genres presents additional concerns that have to be foreseen while processing the data [Sitchinava 2019].

Existing approaches to pos- and morphological tagging texts in pre-modern Russian are effective and even show results that are close to being state-of-the-art for high-resource languages with rich morphology (90% or more than 95%) in the task of pos tagging. Despite that, modern tagging methods are still too sensitive to the genre and origin of the training data. In this paper, we investigate to what

extent the homogeneity and amount of training data affect NLP results on historical texts. A set of experiments was conducted based on the Middle Russian corpus using UDpipe [Straka, Straková 2017], a pipeline that can be trained for tokenization, tagging, lemmatization, and dependency parsing. We focus on the accuracy of pos-tagging and full morphology tagging depending on the structure of the training data.

The paper is structured as follows: Section *Background* outlines the state-of-the-art approaches to full morphological tagging for historical data in pre-modern Slavic. Section *Data* introduces the corpus we work with and some of its essential properties such as size, format, heterogeneity of the samples. Section *Method* is devoted to the step-by-step experiment structure and clarifies the approaches and tools that we used. The most common errors, possible reasons for their appearance, and the overall results are discussed in the sections *Results*, *Analysis*, and *Conclusion*.

2. Background

There are different approaches in NLP to perform pos-tagging and morphological annotation on pre-modern Slavic texts: from sophisticated rule-based systems powered by grammatical dictionaries [Baranov 2007] to labeling by precedent and tagging projection from Modern Russian [Mishina 2016]. The two open tools for processing texts in Middle Russian were developed for the RNC and TOROT copora. The RNC analyzer [Gavrilova et al. 2016] is ruled-based and was developed for annotating the RNC subcorpus of historical texts. TOROT [Berdicevskis et al. 2016] follows a different approach to annotation, a statistical one, and has a distinct advantage over the RNC analyzer disambiguating in cases when several interpretations are possible. These analyzers use two different tagsets which comlicates the comparison of pos-tagging accuracy. For this reason, RNC Middle Russian and TOROT were converted to the UD standard [Zeman et al. 2019]. [Lyashevskaya 2019] introduces a mapping between the RNC and UD tagsets and details the necessary guidelines for full morphological tagging in both schemas. The harmonization of data annotation becomes an essential step for improving the quality of tagging.

[Berdicevskis et al. 2016] combine the RNC (rule-based) and the TOROT (statistical) approaches. The TOROT analyzer surpasses the RNC tool in pos-tagging because the statistical model does not refer to lemmatization. This step is quite complicated to perform because of the unnormalized nature of the data. However, these two methods may complement each other in the task of full morphological annotation. In essence, the RNC model copes better with lemmatization, while the statistical tool performs pos and morphological tagging with higher accuracy.

Since Middle Russian is a highly inflected language, the challenging step of normalization was significant in processing the historical data. However, modern neural network taggers such as MarMot and CLSTM show the state-of-the-art results between 90% and 95% without normalization, which makes pos-tagging easier in terms of practical application [Scherrer et al. 2018]. Despite these results, further experiments revealed that modern tools are depend on genre, origin, and other properties of the training data.

3. Data

For this study, we conducted a set of experiments on the material of the Middle Russian subcorpus of RNC included in the collection of historical corpora. This section reports on the structure of the corpus, including the distribution of texts and tokens over time, the distribution of language register and presents the annotation scheme.

3.1. Data size

The Middle Russian corpus consists of 5,673 historical documents from 1,125 to 1,731. Distribution of documents by time is shown in **Figure 1** (see also **Appendix A**). Most of the texts were created (or copied) from the 15th to the 17th cc.

The corpus contains 8,359,782 tokens. Most of the corpus volume belongs to the 16th-17th cc., see **Figure 2**. The subcorpora of the 16th and 17th cc. are almost equal in terms of tokens, but the 17th c. subcorpus contains shorter documents such as business letters and thus includes more documents than the 16th c. subcorpus.

3.2. Register variation

The corpus texts belong to a large variety of genres. For our study, we reduced them to one of the following language register types: official, business, colloquial, Church Slavonic, and hybrid (e.g., a combination of Church Slavonic and colloquial speech in one document). Some texts were assigned to more than one type (mixed). **Table 3** shows the text and tokens distribution by register. Both distributions are biased toward business and hybrid.

The presence of texts with different language variation labels in our corpus presents another source of difficulties for pos-tagging and tagging of morphological features.

Each variation has its characteristics concerning spelling, grammar, and lexicon. Church Slavonic texts especially stand out, representing essentially not just a language variation, but a separate language. All these discrepancies among texts have a negative impact on model performance. See also Appendix C for the distribution of tokens by the century of creation and register variation.

3.3. Annotation scheme

All documents were segmented and annotated for parts of speech and morphological features according to the UD schema. Sentences are presented in the CoNLL-U format. Each line represents an annotated token with tags separated by single tab characters. We use a standard set of 17 universal pos-tags, and 15 morphological features, each of them takes one to eight possible values. Full overview of the UD tagset is presented in official UD documentation.

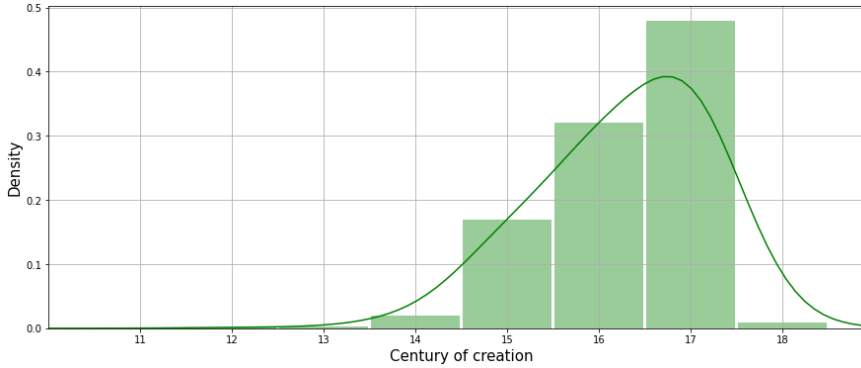

Figure 1: Distribution of documents by year of creation

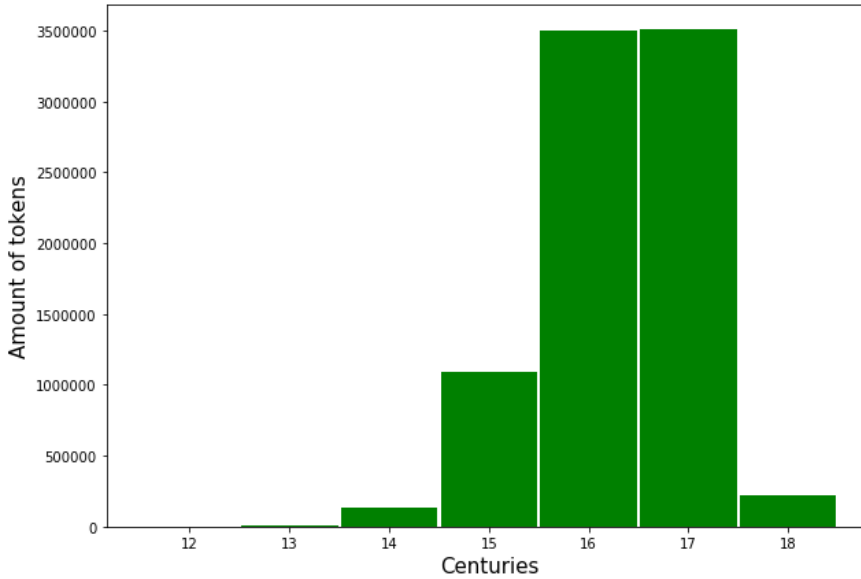

Figure 2: Distribution of tokens by century

Table 1: Distribution of documents by language register

language variety	amount of documents	amount of documents, %	amount of tokens	amount of tokens, %
business	4,718	83.165873	4,131,001	48.840139
hybrid	436	7.685528	3,529,656	41.730537
colloquial	350	6.169575	79,795	0.943403
church	89	1.568835	247,871	2.930538
official	65	1.145778	88,740	1.049158
mixed	15	0.264411	381,146	4.506225

4. Method

4.1. Tools

We used UDPipe, a trainable pipeline designed for tokenization, tagging, lemmatization, and dependency parsing of CoNLL-U files. The version we are using is 1.2.1. The default settings, concentrating on the influence of training data changes on the accuracy of tagging were used.

4.2. Experiment design

In order to evaluate tagging accuracy(by accuracy we mean micro-F1, which is often used for imbalanced data), we divided our data into test and training. For the test part, we randomly choose 10% of the data (678 documents, 846,041 tokens). The remaining part was used for training.

We extracted text samples in three different ways to prove three hypotheses.

Our first goal was to explore whether the training size affects the accuracy of the model. The obvious hypothesis was that the larger the training size, the better. Five samples were extracted from the original corpus, containing 10%, 25%, 50%, 75%, and 100% of the training dataset. We trained five models on these samples and compared the accuracy of pos-tagging, feature tagging, and the overall accuracy.

Our second goal was to analyze the effect of time homogeneity on the model performance. The hypothesis was that a model trained on text from a particular period is better suited for parsing text from the same time. However, the question remains—which part of the corpus is better for training the 'universal' model? Some grammatical forms and additional features are used only in particular periods. Some of them are gradually disappearing, and some new also emerge (for example, aorist tense forms inherent to some verbs like 'бысть' or 'иде' prevail in the 15th c. whereas the auxiliary 'бы' is used more in the 17th c.). To prove our second hypothesis, we took three samples that contain texts from the 15th, 16th, and 17th cc. The sample sizes are 948,678 tokens in the 15th century-sample, 3,312,715 tokens in the 16th century-sample and 3,146,949 tokens in the 17th century-sample.

One more way of sampling is to normalize by sample size, which could be considered as part of the second hypothesis. However, its quantitative characteristics can also be used as an argument for our first hypothesis. This sample contains texts from the 15th, 16th, and 17th cc.,the sample size is one-third data for the 16th or 17th cc. The model trained on this sample should approve or disprove the efficiency of the 'middle way'-approach based on an intuitive view on historical language non-homogeneity—to train a universal model, a universal sample containing every possible deviation from the 'prototype' should be used. However, as we see later, such an approach is flawed and cannot be used for our particular task.

Figure 3 represents the percentage of pos categories in the normalized samples. The distribution differs in all three periods. Verbs and nouns undergo the most significant change over time and we assume that this has an impact on the tagging accuracy.

To reduce the possibility of errors during preprocessing and creation of training sets we replicated our experiment using the updated version of the corpus and new training sets. The results were the same which means errors listed above now seem to be less possible.

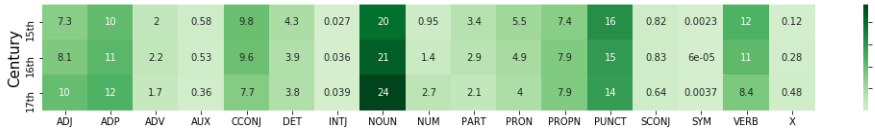

Figure 3: Distribution of part of speech categories in the normalized samples

5. Results

5.1. Effect of training data size variations

The purpose of our first experiment was to see how changes in training data size affect tagging accuracy. The results of this experiment are presented in Figure 4. We consistently reduced the size of the training set from 100% to 10%, and accuracy gradually improved from 60.72% to 93.05% for pos-tags and from 47.21% to 86.06% for morphological tags (see Appendix D for full details).

The smallest training dataset size gave leads the best to a better result (10% model).

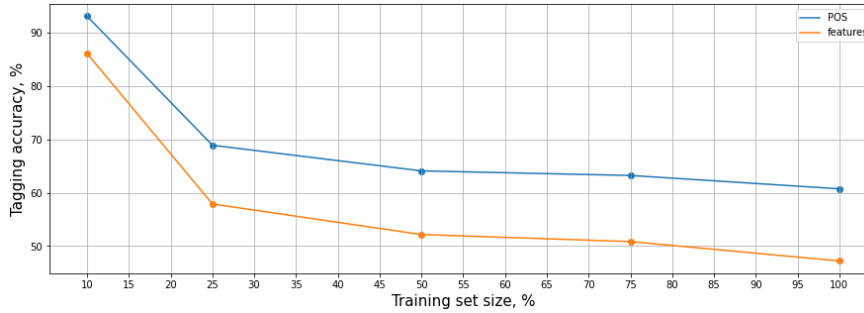

Figure 4: Size dependence of the accuracy

5.2. Time homogeneity effect

In order to find out which part of the corpus is better for training the universal model, we trained three different models. The results are in Figure 5. The 15th-century model demonstrated the best result compared to other models.

5.3. Effect of size-balanced data

By comparing the accuracy of the models trained on all the documents from the 15th, 16th, and 17th cc. and the balanced version with equal-sized chunks from all the periods, we can assume that balancing affects the overall quality. However, such an effect is not significant: 6.5% for pos-tagging and 6% for feature tagging. The results are presented in Figure 6.

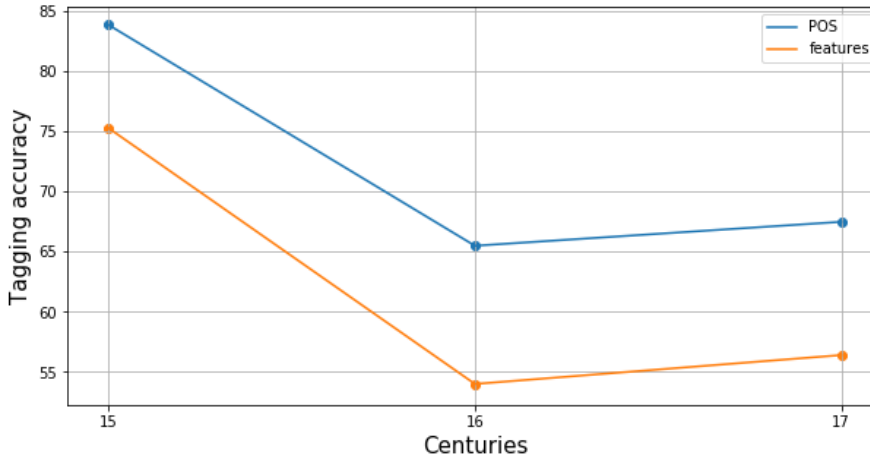

Figure 5: Data split by centuries

Table 2: Data by century

Training set	POS tags accuracy	Morphological features accuracy
15 century	83.82	75.24
16 century	65.46	53.97
17 century	67.45	56.37

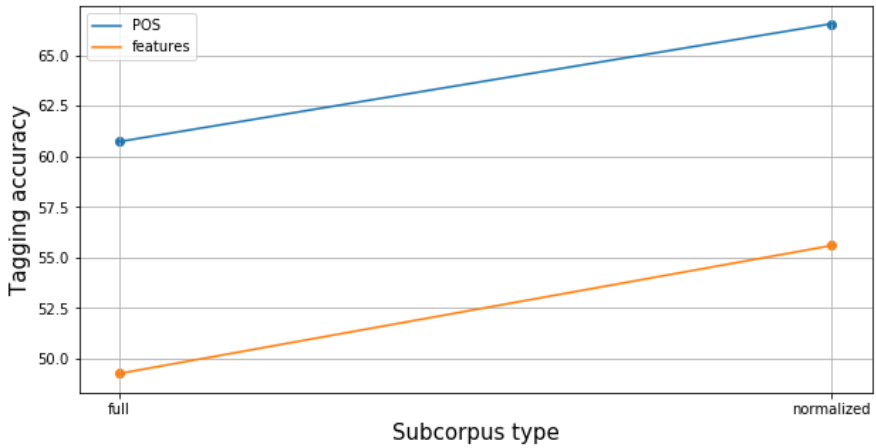

Figure 6: Effect of size-balanced data on tagging accuracy

6. Analysis

Our study showed that bigger does not always mean better. The model trained on 100% of the training dataset showed the worst results while the model trained in 10% of the dataset showed the best results. This leads us to the conclusion that the general problem is overfitting.

The universal model is not always the best solution. The model trained on century-normalized text extracted from all three centuries demonstrated poorer results than the 17th and 15th cc. models.

Despite the 'winners' being obvious, we compared less accurate models to analyze some common mistakes inherent to all of them. This analysis allows us to list some of the most common problems and continue working on improvement strategies.

6.1. POS tagging

To visualize the most common errors, we calculated confusion matrices for pos-tags. **Figures 9–12** show that the main problem is tagging short words from closed classes which include determiners, particles, conjunctions, adpositions, auxiliary verbs, and pronouns. The models we trained confused these parts of speech with verbs and adjectives. The proportion of such confusion differs from model to model. As we see later, this can be explained by the number of particular pos-tags in different training sets.

Russian has undergone significant changes over time. One example is the gradual decrease of conjunctive verb frequency and even their partial disappearance. Conjunctive verbs are mostly short, and therefore some short words from closed classes can be easily confused with them. Some of our training datasets were century based, therefore we should assume that their distribution can affect the quality of pos-tagging.

A significant number of particles, auxiliary verbs, coordinating conjunctions, and adpositions were tagged as verbs by the 16th cc. model (see **Figure 7**). In contrast, the 17th-century model tagged some of them as adjectives and other parts of speech (see **Figure 8**).

Comparing models trained on the 16th and the 17th century dataset, we also noticed that pos categories most often erroneously tagged as nouns were coordinating conjunctions in the 16th c. and subordinating conjunctions in the 17th c. By looking at examples of errors, we discovered that the 16th-century model tends to predict noun tags for conjunctions "и" and "а." We assume that this is related to the number of abbreviations in the 16th-century dataset. 4,598 abbreviated nouns are only one character in this dataset. In comparison, there are only 115 in the 17th-century dataset. This resulted in the better performance of the model. We presume that the reason for the high rate of errors with subordinating conjunction in the 17th-century model is the small number of subordinating conjunctions in our dataset. The 17th-century dataset contains only 3,916 tokens with this pos-tag, whereas there are 13,137 such tokens in the 16th-century one. The 17th-century dataset does not contain enough examples of subordinating conjunctions, which leads to a higher error rate.

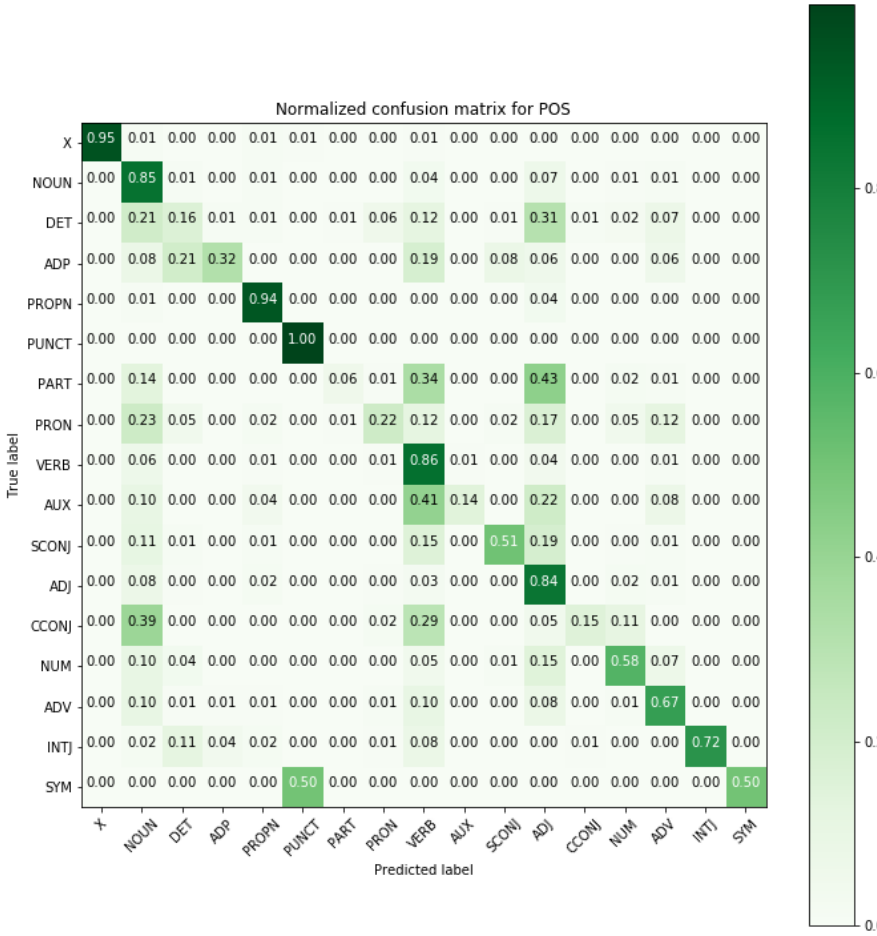

Figure 7: Confusion matrix for model trained on 16 century set

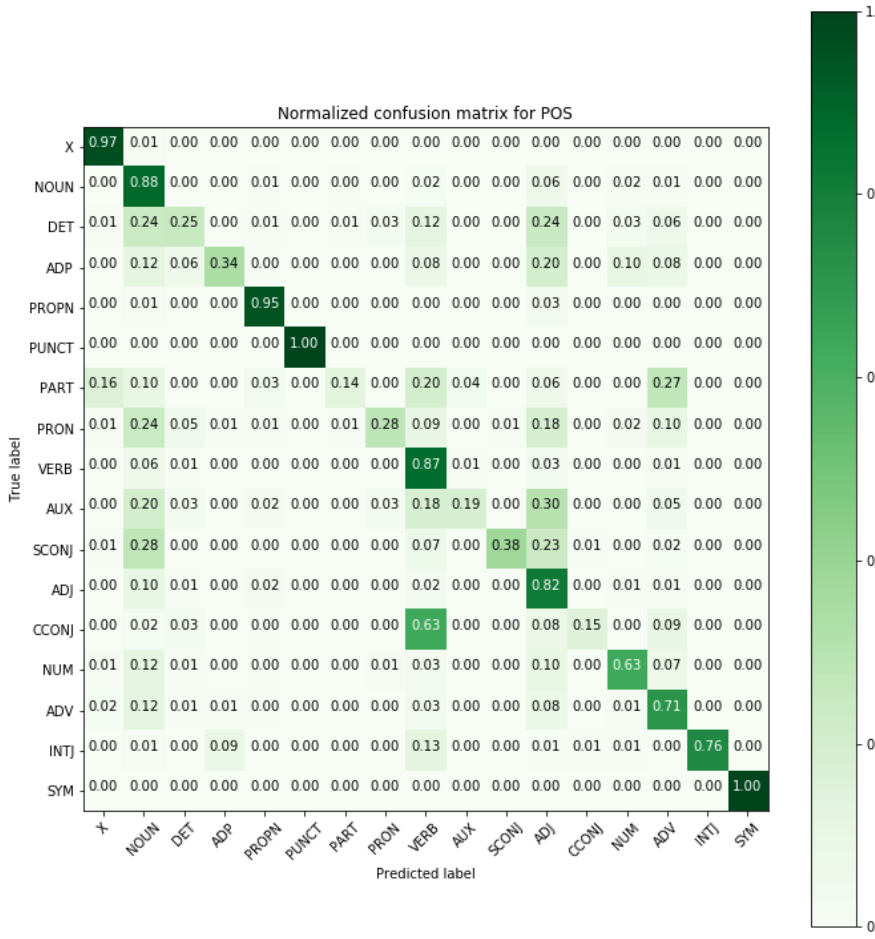

Figure 8: Confusion matrix for model trained on the 17-century set

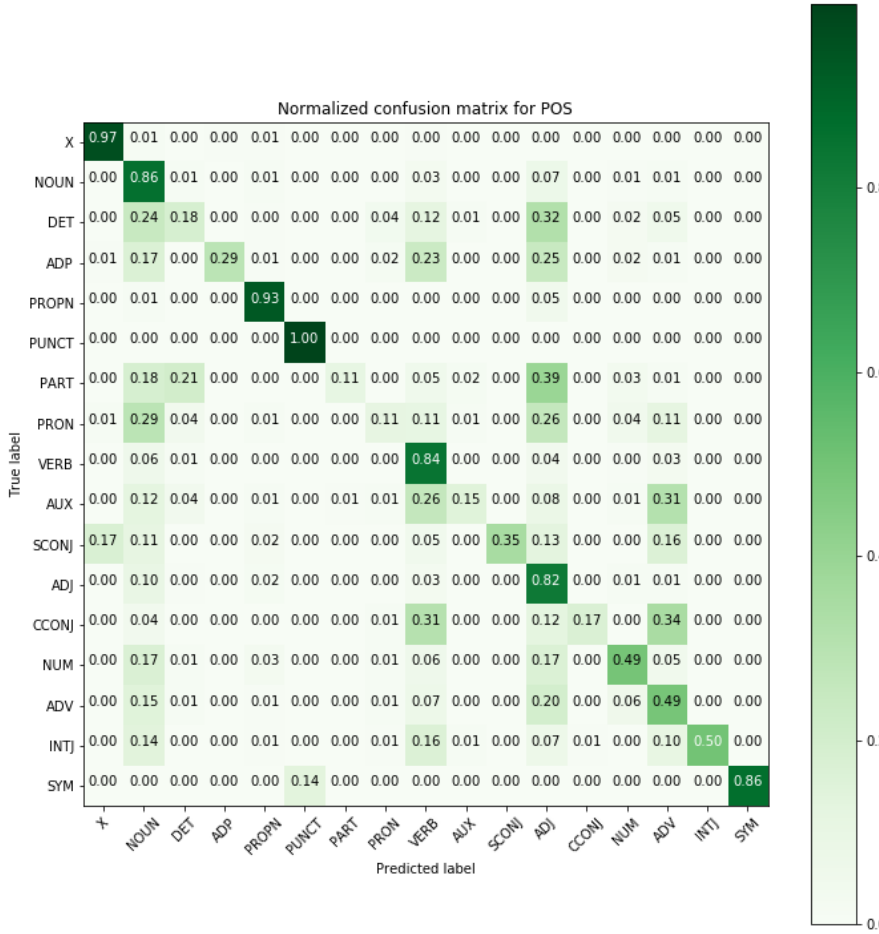

Figure 9: Confusion matrix for model trained on 50%

The ADJ-tag distribution also provided us with considerably impressive results: words from closed classes were frequently confused with adjectives(see. There are a lot of short variants of adjectives throughout the corpus, and their distribution in the training dataset can significantly affect pos-tagging efficiency.

The model trained on 50% of the dataset tagged a lot of determiners, adpositions, particles, and pronouns as adjectives (see Figure 9). The 25%-model showed a different distribution of errors (see Figure 10). If we look closer at 3-letter adjectives, a significant difference in numbers appears. There are almost four times more 3-letter adjectives in the 50%-dataset than in the 25%-dataset. Such a difference in adjectives distribution can explain the short words tagging errors made by the 50%-model.

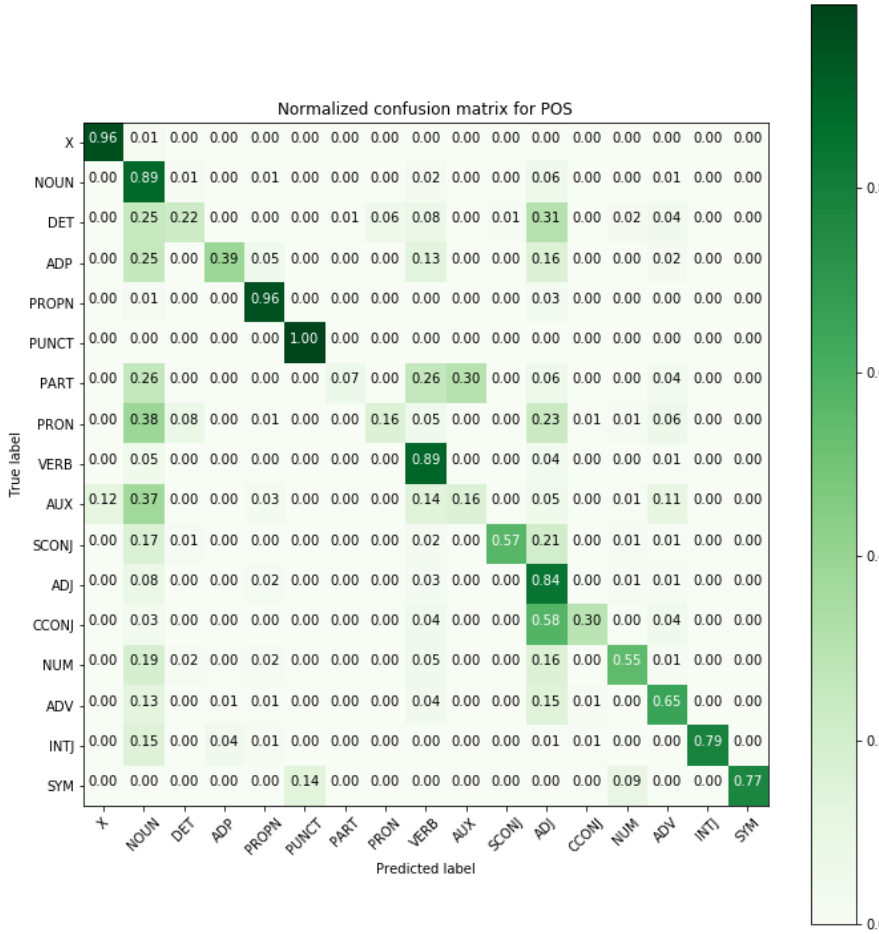

Figure 10: Confusion matrix for model trained on 25%

6.2. Morphological tagging

Results for feature tagging differ from model to model. Some features such as noun case, verb form, tense, voice, mood, gender for verbs, and adjective case are tagged correctly. However, there are also a lot of poorly tagged features. Number and animacy for adjectives, number, abbreviation, and animacy for nouns are the worst. The most frequent error in number tagging is confusing dual with singular or plural. We consider that the main reason for such errors is the similar endings of dual nouns and plural nouns. Some grammatical indicators can have several grammatical functions (grammar syncretism).

Abbreviation tagging accuracy presumably depends on the number of abbreviations in the training data. There is a considerable gap between abbreviation tagging

accuracy for the 16th cc. model, which had a higher proportion of abbreviated nouns in the training dataset and the 17th cc. model. A higher number of abbreviated forms in the training dataset could lead to another problem, discussed above (see tagging short words from closed classes).

Every model had a severe problem with tagging animacy features. Animacy as a morphological feature doesn't have a formal indicator and nouns are inanimate. This is a main reason why about 75% of animate nouns were tagged as inanimate by all models.

6.3. Post hoc analysis

To identify whether the controversial training size effect on the model performance was caused by data bias, we randomly selected 5 more training sets 6.5% of corpus each. If our assumption is correct, we get similar results for each run. As shown in Figure 11, the gap between the 'best' and 'worst' try is about 1%. This proves our initial hypothesis and shows that the training data size can negatively correlate with the model efficiency.

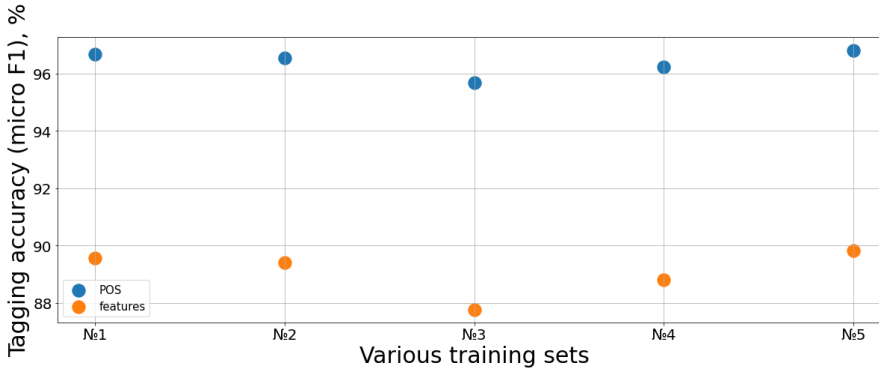

Figure 11: Various training sets of the same size

We also examined how different models tag in-vocabulary (IV) and unseen, out of vocabulary (OOV) words. Here we use the term line in the meaning of a lexeme with its pos- and morphological features, and a word in the meaning of a word form. In Figure 12, Correct IV full tag shows the accuracy of tagging IV-words, and Correct OOV full tag shows the accuracy of tagging OOV-words. Two other metrics, Correct IV unique tokens and Correct OOV unique tokens show accuracy of unique IV-word and OOV-word pos-tagging.

These results highlight two main trends in our models. The models trained on larger sets suit better for tagging OOV-words (85% Vs. 88%), but often cannot tag IV-words correctly. On the opposite, smaller training size models show remarkable results on IV-words but often unable to tag OOV-words correctly.

Taking into account that there are 497,221 unique tokens throughout the corpus, we can now see where this significant gap in accuracy came from. Variations in orthography and grammar create a situation when the borders of the classes become

blurred. It affects some of the most frequent pos-tags like verbs, adjectives, and nouns, so larger models become strongly dependable on their distribution in the training set. As the training set volume increases, a model is trained on more non-standard forms of verbs, adjectives, and nouns. It makes a model more flexible but less accurate. This is the reason why such models are good at tagging unknown words and, at the same time, overfits on frequently used words.

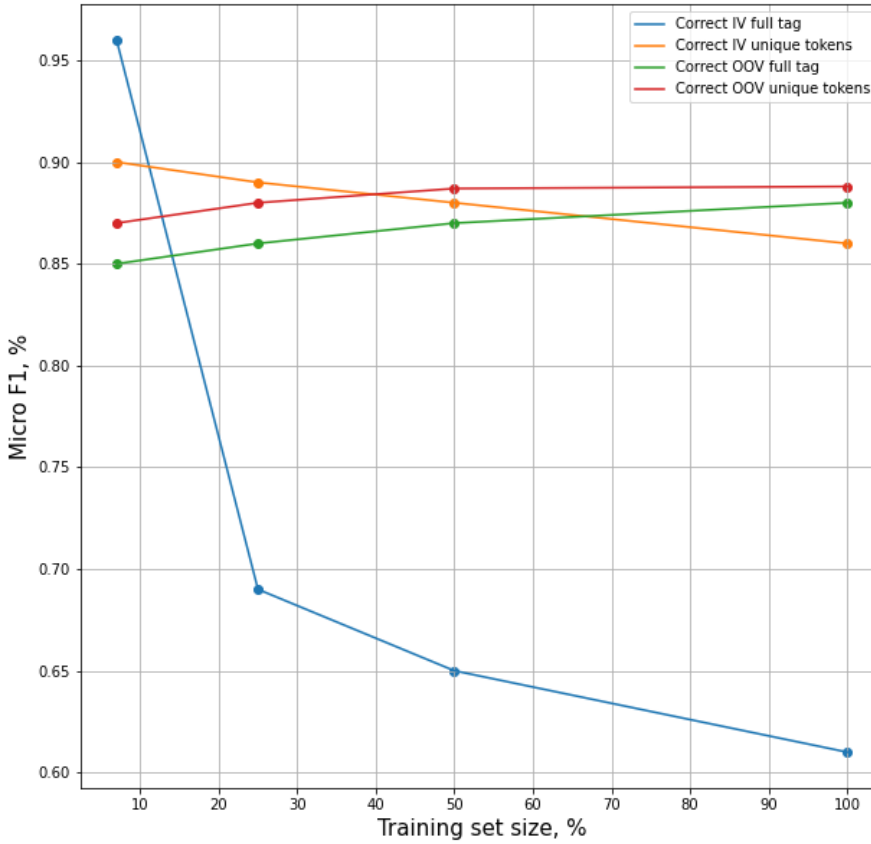

Figure 12: IV and OOV words tagging and training size

The heterogeneity effect can be a possible explanation for our results. We created three training sets (ca. 0.8M, 1.5M and 2.4M tokens) containing only the 17th c. documents of business register to test the classifier on less heterogeneous data. The results are presented in **Figure 13**.

Although with the increase of the training size the model performance eventually drops, the knee point is observed not before the size of 1.5M tokens, whereas in the main experiment, a clear decline in the quality occurs after 600,000 tokens (see **Figure 4**). Thus heterogeneity can account for partial overfitting observed in models trained on samples of different training size.

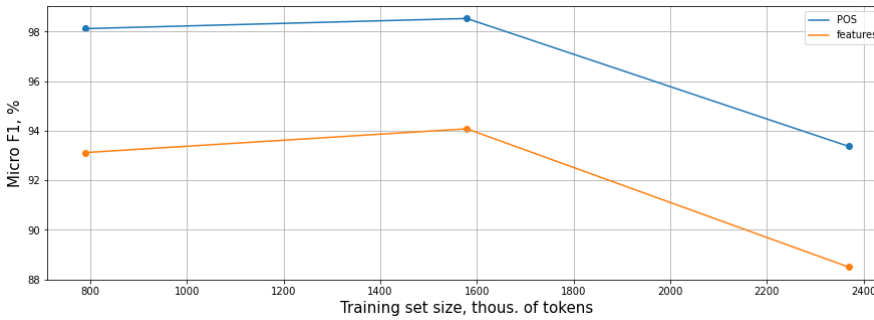

Figure 13: 17th c training sets

7. Conclusions

We trained several models on datasets extracted from the Middle Russian corpus—containing texts from different periods and of different sizes—using UDPipe with default parameters. Datasets were divided into two groups to check how the number of tokens in the dataset can affect overall quality and to understand the possibility of creating a universal model well suited to work with language variety.

Although pos-tagging accuracy was moderate, the results seem to be promising. In future research, we plan to find the middle way between underfitting and overfitting. The fact that models trained on smaller datasets demonstrated better results shows that there may be an optimal dataset-size and model of text sampling.

We identified a probable cause of errors in the tagging of words belonging to closed classes. We plan to use the rule-based approach to solve this particular problem.

The main problem is still heterogeneity. We tried to avoid heterogeneity effect by reproducing the experiment on homogeneous training sets (texts from 17th c. within one particular thematic domain). The results approved our hypothesis: the gap in accuracy between large and small models became less significant (<5% vs. 30%), which means reducing heterogeneity leads to better results. However, this approach cannot be scaled up. Heterogeneity is one of the most important features of historical text, and it is impossible to exclude it from the data artificially.

After the more precise analysis of our models' results, we found out that the models trained on larger datasets cope with OOV-words better. In contrast, the models trained on small datasets showed reliable results for IV-words. A possible solution is to create a hybrid system combining two models oriented on IV or OOV distribution in the training data.

Another improvement scenario is to concentrate on the heterogeneity problem, which could greatly increase the model's sustainability to OOV-words. One possible solution is to use symbol embeddings (like fasttext). This could help us to link some of the OOV-words with their "prototype", which could be useful if we assume that OOV contains different orthography variations of words seen in training data.

8. Acknowledgements

Authors are grateful to

- corpus annotators;
- anonymous reviewers;
- participants of HSE research seminar for fruitful discussions.

References

1. *Arkhangelsky, T., Mishina, E., Pichkhadze, A.* (2014). A System for Digital Morphological Tagging for Old Russian and Church Slavonic Texts. Palaeobulgarica, pp. 21–37.
2. *Baranov V. A., Mironov A. N., Lapin A. N. et al.* (2007), Automatic morphological analyzer of Old Russian language: linguistic and technological solutions [Avtomaticheskij morfologicheskij analizator drevnerusskogo jazyka: lingvistichekie i tekhnologicheskie reshenija] 10th jubilee international conference EVA 2007, Moscow.
3. *Berdičevskis A., Eckhoff H. M., Gavrilova T.* (2016), The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016", Moscow, pp. 99–111.
4. *Gavrilova, T., Shalganova, T., Lyashevskaia, O.* (2017) Processing Orthographic Variation in Lexico-Grammatical Annotation of the Middle Russian Corpus of 15–17th Centuries. By St. Tikhon's Orthodox University: in St. Tikhons' University Review.
5. *Lyashevskaya O.* (2019), A reusable tagset for the morphologically rich language in change: a case of Middle Russian [Mnogocelevoj morfologicheskij standart razmetki dlja jazyka s menjaushejsja grammaticheskoj strukturoj: sluchaj starorusskogo korpusa]Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2019", Moscow, pp. 99–111.
6. *Scherrer, Y., Rabus, A., Mocken, S.* (2018). New Developments in Tagging Premodern Orthodox Slavic Texts. Scripta & e-Scripta, 18, 9–33.
7. *Sitchinava D. V.* (2019) Spelling variation and word clusters in the Middle Russian Corpus. Paper presented at the International Conference Historical Corpora and Variation. Cagliari, 4–5th April 2019.
8. *Straka M., 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 2017 Aug, pp. 88–99.
9. *Zeman, D., Nivre, J., Abrams, M., et al.* (2019). Universal Dependencies 2.5, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-310.

Appendix A

Table 3: Distribution of documents by century of creation

century	number of texts	%
12	9	0.16
13	12	0.21
14	108	1.90
15	962	16.96
16	1,817	32.03
17	2,717	47.89
18	48	0.85

Appendix B

Table 4: Distribution of tokens by century of creation

century	tokens	%
12	1,824	0.02
13	4,333	0.05
14	131,216	1.55
15	1,091,367	12.90
16	3,502,392	41.41
17	3,504,853	41.44
18	222,224	2.63

Appendix C

Table 5: Distribution of tokens by century of creation and register variation

century	business	church	colloquial	hybrid	mixed	official
12	1,824	0	0	0	0	0
13	3,316	0	0	1,017	0	0
14	23,192	2,935	0	105,089	0	0
15	366,584	17,362	0	618,250	431	88,740
16	1,342,611	221,822	0	1,877,796	60,163	0
17	2,319,755	5,752	73,126	785,668	320,552	0
18	73,719	0	6,669	141,836	0	0

Appendix D

Table 6: Variations in training data size

Size of the training set	POS tags accuracy	Morphological features accuracy
10%	93.05	86.06
25%	68.88	57.88
50%	64.08	51.56
75%	63.22	50.84
100%	60.72	47.02