Source: Scripta & e-Scripta

Scripta & e-Scripta

Location: Bulgaria

Author(s): Juliane Besters-Dilger

Title: Neural Morphological Tagging for Slavic: Strengths and Weaknesses

Neural Morphological Tagging for Slavic: Strengths and Weaknesses

Issue: 21/2021

Citation style: Juliane Besters-Dilger. "Neural Morphological Tagging for Slavic: Strengths and Weaknesses". Scripta & e-Scripta 21:79-92.

https://www.ceeol.com/search/article-detail?id=994365

SCRIPTA
& e- SCRIPTA

e-SCRIPTA

21 / 2021

Neural Morphological Tagging for Slavic: Strengths and Weaknesses

Juliane Besters-Dilger, Achim Rabus

Abstract: The neural network tagger CLStM has been applied to the Old Russian *Žitie Evfimija Velikogo* (GIM, Chud. 20), a copy of the second half of the 14th century. The strengths of this tagger consist in its ability to automatically annotate an orthographically non-normalized text with dozens of pages within a few minutes, yielding a high accuracy with respect to part of speech and morphological features. Moreover, the tagger is capable of disambiguating case syncretism to a large extent, even in split constructions. Manual correction of the automatic tagging will result in a correctly tagged text considerably faster than when using a rule-based tagger or tagging completely manually. The weaknesses of the CLStM-tagger comprise certain examples of incorrect POS-tagging, sometimes incomplete or incorrect attribution of morphological categories to some parts of speech. Superscript letters and punctuation can pose special problems, normalization of punctuation will achieve better tagging results. The proportion of correct tags is higher when the token has been seen during the training process; unknown words (OOV) show a higher error rate. In the paper, we analyze the strengths and weaknesses of the tagger by providing specific examples. Furthermore, we demonstrate how to use automatically tagged, uncorrected data for quantitative analysis.

Key words: Neural network tagger, POS and full morphology tagging, context sensitivity, punctuation, quantitative analysis

Introduction

Part-of-speech (POS) and full morphology tagging of pre-modern Orthodox Slavic texts is a challenging task, due to the rich morphology of Slavic and its lack of standardization. Different attempts to address this issue have been made such as developing and applying sophisticated rules for morphological analysis (e.g., Baranov et al. 2007), tagging projection from Modern Russian (Meyer 2011), or the combination of a statistical and a rule-based tagger (Berdičevskis et al. 2016). However, the recent boost in applying artificial neural networks for natural language processing (NLP) tasks has resulted in the development and training of a novel tagger for pre-modern Orthodox Slavic based on neural networks. The tagger CLStM (Church Slavonic long-short-term memory, https://github.com/yvesscherrer/lstmtagger), described in Scherrer, Mocken and Rabus 2018, can be characterized as a bidirectional recurrent neural network based on character representations, not word embeddings. Its ability to cope with variation with respect to orthography, morphology and abbreviations and its context sensitivity are relevant features. The latter entails that, as opposed to rule-based taggers, it automatically disambiguates homonymous forms. This contribution is devoted to an analysis of clear advantages, but also some weaknesses of the tagger's results.

We used a model trained on the TOROT and PROIEL Old Russian and Old Church Slavonic data tagged by Hanne Eckhoff et al. (see, e.g., Eckhoff and Haug 2012, Eckhoff and Berdičevskis 2015) and consisting of more than 200,000 word tokens; the tagset has been converted to the Universal Dependencies tagset (universaldependencies.org). Then we applied the tagger to the Old Russian *Žitie Evfimija Velikogo* (GIM, Čudov collection 20), a copy of the second half of the 14th century.[1] The date of the translation from Greek – the author is Cyril of Scythopolis, a sixth-century monk (Schwartz 1939) – into Slavic is unknown. The text comprises 66 two-column pages.[2]

[1] Our workflow was as follows: First, we converted the Word file of the text to plain text format. Then we tokenized it using UDPipe (http://lindat.mff.cuni.cz/services/udpipe/run.php). Subsequently, we tagged the tokenized file under Linux using the default settings and the updated pre-modern Slavic model available on https://github.com/yvesscherrer/lstmtagger.

[2] We would like to thank the Vinogradov Russian Language Institute of the Russian Academy of Sciences and especially Aleksandra Duhanina for providing us with a copy of this text in Unicode. The Institute and the Department of Slavonic Studies of Freiburg University cooperate in the project "Digital Paleoslavistics (DigiPalSlav)" supported by the Alexander von Humboldt Foundation within their research group linkage program.

Strengths of the CLSTM tagger

The strengths of the CLStM tagger consist in its ability to automatically annotate an orthographically non-normalized text comprising dozens of pages within a few minutes and to yield a high accuracy with respect to part of speech and morphological features. In the case of the *Žitie Evfimija Velikogo*, the accuracy amounts to 96.2% correct part of speech tags and 92.3% correct morphological features.[3] These are approximate figures, based on the manual analysis of around 10% of the annotated text. Moreover, the tagger is capable of disambiguating case syncretism to a large extent, sometimes even in split constructions and with out-of-vocabulary tokens (OOV).

(1) w̃	ADP	
прⷭ́стыꙗ	ADJ	Case=Gen|Degree=Pos|Gender=Fem|Number=Sing|Strength=Weak OOV
вⷱ́ца	NOUN	Case=**Gen**|Gender=Fem|Number=Sing
мⷬ́рыꙗ	PROPN	Case=**Gen**|Gender=Fem|Number=Sing OOV

вⷱ́ца and мⷬ́рыꙗ are correctly analyzed as Gen. Sg. Fem., although the nominative looks identical.

(2) она	ADJ	Case=Nom|Gender=Masc|Number=**Dual**
[…]		
възⷬратиста	VERBA	spect=Perf|Mood=Ind|Number=**Dual**|Person=3|Tense=Past|VerbForm=Fin|Voice=Act OOV
ⷭ	AUX	Case=Acc|Number=Sing|Person=3
в	ADP	_
домъ	NOUN	Case=Acc|Gender=Masc|Number=Sing

The verb phrase възⷬратиста ⷭ в домъ follows two lines after она. Nevertheless, она has been recognized as Nom. Dual Masc. and not Nom. Sg. Fem. or Nom. Pl. Neut. This is quite an impressive achievement and shows that, given a sufficient amount of training data, neural taggers can achieve accuracies that taggers based on other technologies cannot.

[3] In Scherrer, Mocken and Rabus 2018: 16-17 it is explained how to evaluate the accuracy of morphological features (micro-F1).

Weaknesses of the CLSTM tagger and analysis of errors and not yet satisfactory results[4]

However, there are some undeniable weaknesses of the CLStM tagger. As philologists, we are interested especially in the following question: If this tagger has learned a certain context sensitivity, what errors occur and why? In the text under investigation, one can distinguish seven types of errors or unsatisfactory results (see under c and e).

a) There are annotations that are morphologically possible, but do not correspond to the context.

(3) въ ADP _
пропастєхъ NOUN Case=Loc|Gender=Fem|Number=Plur
зємнъыхъ ADJ Case=**Gen**|Degree=Pos|Gender=Fem|Number=Plur|Strength=Weak OOV

Here the context sensitivity is apparently not sufficient for excluding the case genitive. One could suppose that the postposition of the adjective motivates the case error, but in other passages postposition does not prevent correct annotation.

b) There are annotations which are morphologically impossible.

(4) подъимъ VERB Mood=Ind|Number=**Plur**|Person=**1**|Tense=**Pres**|VerbForm=Fin|Voice=Act OOV

This cannot be the 1st person plural of the present, but is the Nom. Sing. of an act. past participle. Nevertheless, one has to acknowledge that if a verb подъити (following the type хвалити) existed, the attribution of morphological features would be correct.

(5) измѣнєныа NOUN Case=**Acc**|Gender=Neut|Number=Sing OOV

измѣнєныа cannot be an Acc. Sing. Neut. We found also a noun разумые annotated as Gen. Sing. Neut.

[4] After completion of our analysis, Yves Scherrer published an improved tagger based on the stanza toolkit: https://github.com/yvesscherrer/stanzatagger. Trained with the same training data, this new tagger performs significantly better than the CLStM tagger: Many of the weaknesses of the CLStM tagger are remedied in the stanza tagger meaning that the stanza tagger tags the majority of the cases discussed here correctly.

The reason for these errors is unclear, because many other nouns on -ныѥ like спасеныѥ, ѣавленыѥ are always annotated correctly. It is notable, though, that in examples (4) and (5) the tokens are marked OOV "out of vocabulary", which means that they have not been part of the training data during model training. земиꙁьıχъ (error type a) belongs to the OOV tokens as well. For further remarks on their susceptibility to errors see below.

c) The attribution of morphological categories is incomplete.

(6) кр̑стивъ VERB Case=Nom|Gender=Masc|Number=Sing|Strength=Strong|Tense=Past|VerbForm=Part OOV
ѥго PRON Case=Gen|Gender=Masc|Number=Sing|Person=3|PronType=Prs
. PUNCT _
и CCONJ _
постриг҃ъ VERB Case=Nom|Gender=Masc|Number=Sing|Strength=Strong|VerbForm=Part|Voice=Act

The features lacking here are for the first verb "Voice=Act", for the second "Tense=Past".

This example is also interesting because the annotation of the pronoun ѥго shows that the category of animacy is not taken into account. Every animate accusative, e.g. б̑а, is analyzed as genitive. This goes back to the training material, the TOROT and PROIEL data, where the category of animacy has been excluded from the morphological tagging, because the morphological annotation "genitive" plus the syntactical annotation "object" have been regarded as sufficient for characterising animate direct objects. But this approach does not seem satisfactory in the case of a purely morphological tagger like CLStM, where the syntactic feature "object" has not been taken into account during model training. The relative majority (13.7%) of the "errors" (or unsatisfactory results) found among the known words concern such "genitives" that are in reality animate accusatives.

d) Incorrect categories are attributed to a part of speech

Examples:

(7) ѥствꙋ NOUN Gender=Fem|**Mood=Ind**|Number=Sing OOV

A noun has no mood.

(8) с̑ѣнаıа ADJ Case=Nom|Gender=Fem|Number=Sing|Strength=Weak|**Tense=Pres** OOV

An adjective has no tense.

(9) знаѥмъ VERB Gender=Masc|Number=Plur|**Person=1**|Tense=Pres|**VerbForm=Part**|Voice=Act OOV

A verb has either a person or is a participle.

The explanation for the error types c and d, i.e. missing, incorrect or superfluous morphological features, is the following: For each token, the tagger predicts all morphological features at the same time and then deletes those which are improbable on the basis of what it has learned. Additionally, the part of speech-tagging and the attribution of morphological features take place simultaneously, so that the part of speech does not predetermine the features. Yves Scherrer has recently trained pre-modern Slavic models with a modified tagger based on the stanza toolkit (https://stanfordnlp.github.io/stanza/index.html). Crucially, stanza first determines the POS and subsequently the morphological features, thus excluding nonsensical combinations of morphological features. Because of that, these errors will probably be a thing of the past soon. The three lexemes in question are marked OOV as well.

e) The attributed part of speech is unsatisfactory.

As shown in example (2), the pronoun ѡна is analyzed as an adjective (ADJ). In other passages, the same is true for the demonstrative pronouns тъ and сь. Of course, онъ, тъ and сь can, if they accompany a noun, function as adjectives and be congruent in case, number and gender (like in, e.g., по ѡноѩ сторонѣ, сию заповѣдь, тоѩ цьрквѧ). In this case, they are analyzed as determiners (DET). But when they stand alone, they function primarily as (substantival) pronouns.

These attributions as adjectives go back to an incomplete transfer from the training material TOROT and PROIEL, where the POS tagging consists always of two elements: In the first column, онъ, тъ and сь are analyzed as adjectives or determiners, but in the second as pronouns (Pd = demonstrative pronoun). The CLStM tagger takes over only the first annotation and neglects the second, so that it has learned an incomplete annotation.

f) Errors are based on abbreviations (rare).

(10) пѧ҃ **VERB** Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act OOV

пѧ҃ is not a 3rd person Sing. Aor. This is probably a wrong analogy to ре҃ which means рєчє.

g) Errors are caused by the fact that superscript letters were put on the line, but without any special marking.

(11) ползу **VERB** Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act

ползу has been – in principle correctly – analyzed as 1st person Sing. Pres. of the verb ползти. But the context and the Greek original make clear that this is the Acc. Sing. Fem. of the noun польза. The *l* is a superscript letter, therefore the yer' disappeared.

In order to minimize this error type, it would be necessary to conduct some preprocessing steps before tagging. Superscript letters should not be simply placed on the line, but be marked, because it is likely that in their environment a reduced or full vowel has disappeared.

Which further pre-processing steps can be conducted to improve the results? One main reason for incorrect tags is punctuation. Since the tagger relies on sentence boundaries, punctuation marks signaling the end of a clause such as full stops that appear in the middle of a pre-modern Slavic clause will inevitably confuse the tagger and lead to an incorrect interpretation of syntactic relations and, thus, morphological tags. Therefore, we tagged the text twice: first in its original version with an enormous amount of superfluous – more or less randomly distributed – full stops, second in a version with corrected punctuation, so that clauses and phrases remained complete. In the second case, we got better tagging results, because contextual relations were better understood.

(12) подасть VERB Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
намъ PRON Case=Dat|Gender=Masc|Number=Plur|Person=1|PronType=Prs
. PUNCT _
взитыѥ NOUN Case=**Nom**|Gender=Neut|Number=Sing OOV

After removal of the full stop, the last token is correctly annotated:

взитыѥ NOUN Case=**Acc**|Gender=Neut|Number=Sing OOV

(13) подасть VERB Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin| Voice=Act

бъ	NOUN	Case=Nom|Gender=Masc|Number=Sing
гр҃ечьстѣи	ADJ	Case=**Gen**|Degree=Pos|Gender=Fem|Number=Sing|Strength=Weak
области	NOUN	Case=**Gen**|Gender=Fem|Number=Sing OOV
.	PUNCT	_
и	CCONJ	_
своимъ	ADJ	Case=Dat|Gender=Fem|Number=Plur|Person=3|Poss=Yes|Reflex=Yes
цр҃квамъ	NOUN	Case=Dat|Gender=Fem|Number=Plur
дара	NOUN	Case=Gen|Gender=Masc|Number=Sing
бнѧ	ADJ	Case=Gen|Degree=Pos|Gender=Masc|Number=Sing|Strength=Strong

After removal of the full stop in the middle of the clause, гр҃ечьстѣи and области are correctly annotated as datives.

Our analysis has shown that the majority of the errors that occurred concern out-of-vocabulary (OOV) tokens. We calculated the error rate of words that had been seen during training in relation to error rates with OOV tokens.[5] The results are as follows:

Table 1: *Error rates with known and unknown word tokens*

	Percentage	Incorrect annotations
Tokens seen during training	63.5%	35.6%
OOV tokens	36.5%	64.4%

As can be seen, OOV tokens amount to roughly one third of the text, but their error rate amounts to two thirds of all errors.

Our analysis of the strengths and weaknesses of the CLStM tagger for pre-modern Slavic has shown that it is a very valuable tool for the fast and highly accurate automatic morphological tagging of texts. In order to achieve a completely correctly tagged text, merely a limited number of manual corrections has to be made. Using the CLStM tagger with subsequent manual post-correction will result in a correctly tagged text considerably faster than when using a rule-based tagger or tagging completely manually. The models we used feature the Universal Dependencies tagset

[5] The following figures do not distinguish between errors concerning the part of speech or one or several morphological features; any kind of deviation has been counted.

which shows a high ratio of equivalent annotations with the tagset of the Russian National Corpus (Lyashevskaya 2019: 11). Conversion routines and scripts already exist meaning that it is possible to employ the CLStM tagger for pre-tagging materials for use in corpora such as the historical parts of the Russian National Corpus.

As main reasons for errors committed by the tagger, punctuation, the treatment of superscript letters in the text, OOV tokens and the category of animacy are to be mentioned. Another source of errors, the combination of morphological features that don't make any sense for the part of speech in question (see above, c and d), will probably soon be eliminated. Due to the 'long tail' of many word tokens that occur extremely infrequently in corpora, increasing the amount of training data will most likely result in only a slight decrease in OOV tokens, especially given the high amount of orthographic variation in real-life pre-modern Slavic data. A more promising approach to improve tagging results is to train new taggers using tools such as stanza that take into account the mutual dependency of POS and morphological features, thus preventing the errors sometimes encountered with CLStM that impossible and non-matching morphological features are assigned to a specific POS.[6] Another step to consider would be to retrain the model with a specific genitive-accusative tag.

Proof of concept – quantitative analysis using uncorrected data

In this section, we report on a pilot study where we use the automatically tagged data without manual post-correction for linguistic analysis. In doing so, we want to demonstrate that structures in historical texts can be uncovered using a quantitative approach without manual post-processing steps. While automatically tagged, uncorrected data for quantitative linguistic analysis has been used for quite some time with respect to English (e.g., de Haan 1997), we are unaware of any studies using this approach for pre-modern Orthodox Slavic.

Taking up on the famous Labovian dictum that historical linguistics is the "art […] to make the best of […] bad data" (Labov 1972: 100), we want to show that in the Digital Age, a large amount of automatically tagged, uncorrected data is by no means worse than the traditional source of historical linguistics, a small amount of data that has been carefully corrected manually. Given the increasing possibilities of AI-assisted Handwritten Text Recognition applications such as Transkribus (Rabus 2019) and neural tagging as demonstrated in the current paper, in the near future,

[6] As mentioned before, a tagger based on the stanza toolkit has been trained after completion of our analysis: https://github.com/yvesscherrer/stanzatagger.

there will be a considerable boost in digitally available historical linguistic data that cannot possibly be manually corrected (let alone tagged) in a reasonable amount of time and for a reasonable amount of money. Because of that, there is need for the historical linguistics community to build comfort with incompleteness and margin of error when it comes to quantitative analysis (see Dombrowski, this volume). There is no evidence to believe that the analysis of a small fraction of available data that has been manually corrected leads to better generalizations than the analysis of a larger portion of data without manual corrections (cp. Piper 2020).

For our quantitative analysis, we used the complete Evfimij text automatically tagged using CLStM and consisting of 16,075 tokens. The analysis was conducted in AntConc (https://www.laurenceanthony.net/software/antconc/), since AntConc allows for rather simple corpus searches that take into account POS and full morphology tags. Nevertheless, some preprocessing steps (search and replace operations) had to be taken in order to make full use of the tags within AntConc.

In order to get an idea of the real-life accuracy and practical usability of the tagger, we evaluated the amount of false positive and false negative results when searching for surface strings on the one hand and for tags assigned by CLStM on the other. We searched for all word tokens ending in -ша, the typical ending of 3rd person plural aorist, and compared the results to the search of tokens tagged with aorist (i.e. perfective past) tokens in the third person plural. Searching for -ша yielded 140 results, while 99 tokens were retrieved searching for the 3rd person plural aorist tag.

Among the numerous false positive results of the surface search (tokens ending in -ша), many were past participle forms such as ѡбѣцıавша, ѿшедша, оумеръша. These forms had been annotated with other tags and were, thus, correctly ignored when searching for the aorist tags. The only false positive past participle present in both search results is оусѡпша, incorrectly tagged as an aorist form.

Besides the past participles, the surface search yielded several other false positives, among them the pronoun вша and the noun оужоша. The only false negative result of the tag search, i.e. a correct aorist form that could not be found when searching for tags (due to its being tagged differently), but only by using the surface search, is ѣша. This form was incorrectly tagged as a participle form. Overall however, it is obvious that searching for tags yields incomparably more correct results than searching for surface forms as well as only a very small number of false negatives. Because of that, quantitative analysis based on uncorrected tagging results can be conducted with a tolerable margin of error.

Our pilot study is concerned with the distribution of finite past tense forms. In the Universal Dependencies tagset, imperfects are marked with Aspect=Imp and Tense=Past, whereas aorists are marked with Aspect=Perf and Tense=Past. Using AntConc and searching for the respective tags, we found a total of 1,035 tokens

tagged as either imperfects or aorists. While there is a 'long tail' of tokens appearing just once or twice, there are several forms that appear more often. The most frequent tokens are the following:

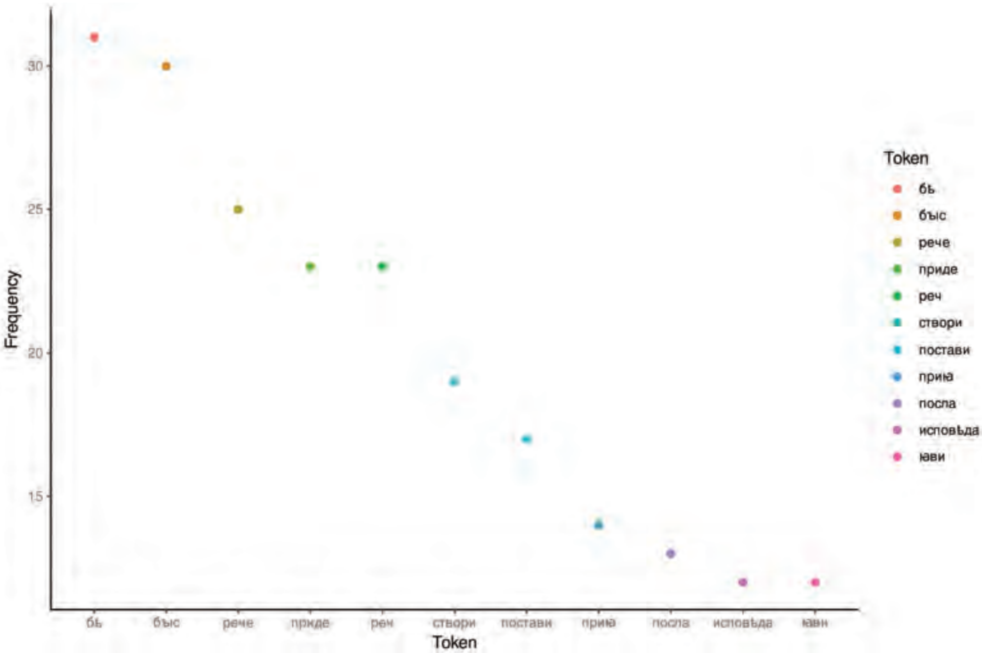

Figure 1: *Most frequent imperfect and aorist tokens*

As can be seen, 100% of the most frequent tokens are in the 3rd person singular, which was to be expected, given that the text in question is devoted to narrating the life of a saint. Crucially, since neither orthographic normalization nor lemmatization were conducted prior to analysis, рєчє and рєⷱ (with unmarked superscript ⷱ) appear as two different forms.

With respect to the overall distribution of imperfect and aorist forms (Figure 2), we see once again that singular forms are considerably more frequent than plural forms, both in the imperfect and in the aorist. Moreover, we see that aorist forms (marked in light color) occur considerably more frequently than imperfect forms (marked in dark color):

When comparing the relation of aorist vs. imperfect forms in Evfimij (806 vs. 229 tokens) to the manually tagged (resp. corrected) Lives of Sergij of Radonež and Stefan of Perm as present in the TOROT corpus (nestor.uit.no), one encounters a statistically significant difference to the situation in the Life of Sergij (820 vs. 463 tokens), while the difference to the Life of Stefan (524 vs. 143 tokens) is not significant. Duhanina

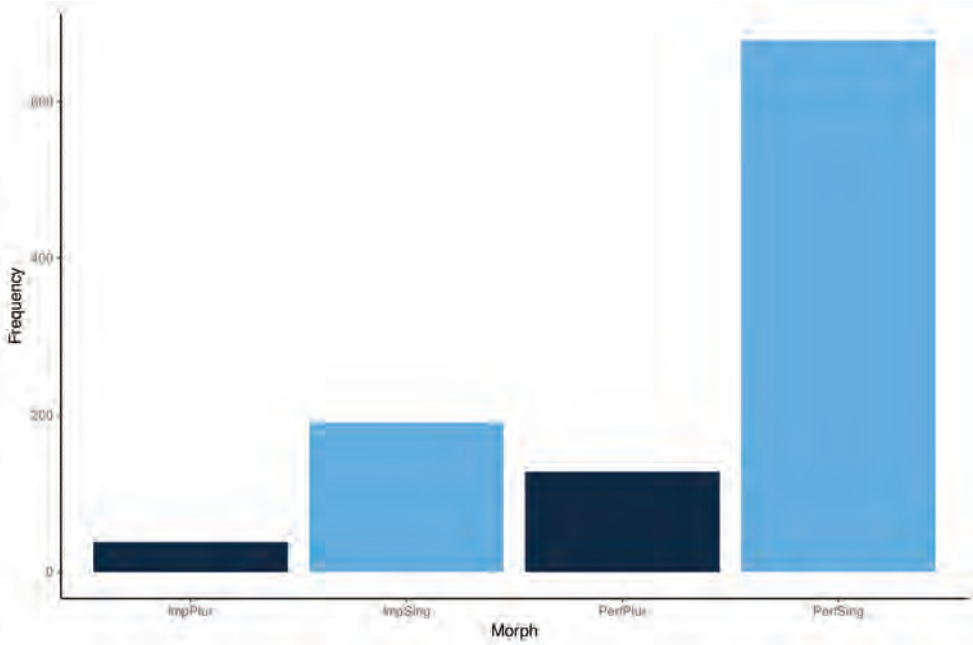

Figure 2: *Frequency of aorist and imperfect forms*

(2008) has shown that the relative frequency of imperfect forms in comparison to aorist forms is considerably higher in the Life of Sergij than in the Life of Stefan, suggesting that the relation found in the Lives of Evfimij and Stefan is the unmarked one for this genre. Undoubtedly, this issue needs further investigation.

While the actual added value of the linguistic analysis conducted in this pilot study should not be overestimated, we hope to have given an outlook on what kind of linguistic analysis can be done with automatically tagged texts even without manual post-correction.

Conclusion

In this contribution we have shown that, despite some weaknesses, current neural network taggers for morphologically rich languages such as varieties of pre-modern Slavic perform well enough to make quantitative linguistic analysis possible even without manual post-correction of tags. While a certain amount of noise due to erroneous tags has to be taken into account – which will likely get smaller as the neural taggers will continuously improve – these taggers mark an important step for the quantitative analysis of pre-modern Slavic texts. The

need to develop and improve workflows for quantitative linguistic analysis that require only minimal manual labor will inevitably rise given the steady increase of both historical corpora and Handwritten Text Recognition models and tools such as Transkribus. Neural taggers such as CLStM are crucial elements in that process.

REFERENCES

Baranov et al. 2007: Баранов, В. А., А. Н. Миронов, А. Н. Лапин, И. С. Мельникова, А. А. Соколова, Е. А. Корепанова. "Автоматический морфологический анализатор древнерусского языка: лингвистические и технологические решения." В *10-я юбилейная международная конференция «EVA 2007 Москва».* Москва, 2007 – URL: http://conf.evarussia.ru/upload/eva2007/reports/doklad_1318.pdf <accessed 08.08.2021>.

Berdičevskis et al. 2016: Berdičevskis, Aleksandrs, Hanne Eckhoff, Tatiana Gavrilova. "The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian." In *Proceedings of the Annual International Conference "Dialogue"* (2016). Ed. V. P. Selegej et al. Moskva: Izdatel'stvo RGGU, 2016, 99–111. (Computational Linguistics and Intellectual Technologies 15 (22)).

de Haan 1997: de Haan, Pieter. "An experiment in English learner data analysis". In *Studies in English language and teaching*. Eds. J. Aarts, I. de Mönnink, H. Wekker. Amsterdam – Atlanta, GA: Rodopi, 1997, 215–229.

Duhanina 2008: Духанина, Александра В. *Морфологические нормы в сочинениях Епифания Премудрого (система глагола)*. Автореферат диссертации. Москва, 2008.

Eckhoff and Berdičevskis 2015: Eckhoff, Hanne, Aleksandrs Berdičevskis. "Linguistics vs. Digital Editions: The Tromsø Old Russian and OCS Treebank." *Scripta & e-Scripta* 14–15, 2015, 9–25.

Eckhoff and Haug 2012: Eckhoff, Hanne, Dag Haug. "The PROIEL Corpus as a Source to Old Church Slavic: A Practical Introduction." *Преславска книжовна школа* 12, 2012, 368–383.

Labov 1972: Labov, William. "Some Principles of Linguistic Methodology." *Language in Society* 1/1, 1972, 97–120.

Lyashevskaya 2019: Lyashevskaya, Olga N. "A Reusable Tagset for the Morphologically Rich Language in Change: A Case of Middle Russian. In *Proceedings of the International Conference 'Dialogue' (2019)*. Ed. V. P. Selegej et al. Moskva: Izdatel'skij centr "Rossijskij gosudarstvennyj gumanitarnyj universitet", 2019, 1–14 (Computational Linguistics and Intellectual Technologies 18 (25)).

Meyer 2011: Meyer, Roland. "New wine in old wineskins? Tagging Old Russian via annotation projection from modern translations." *Russian Linguistics* 35/2 (2011): 267–281.

Piper 2020: Piper, Andrew. *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*. Cambridge: Cambridge University Press, 2020.

Rabus 2019: Rabus, Achim. “Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus.” *Scripta & e-Scripta* 19, 2019, 9–32.

Scherrer, Mocken and Rabus 2018: Scherrer, Yves, Susanne Mocken, Achim Rabus. “New Developments in Tagging Pre-Modern Orthodox Slavic Texts.” *Scripta & e-Scripta* 18, 2018, 9–33.

Schwartz 1939: Schwartz, Eduard. *Kyrillos von Skythopolis*. Leipzig: J.C. Hinrichs, 1939.

[Baranov, V. A., A. N. Mironov, A. N. Lapin, I. S. Mel’nikova, A. A. Sokolova, E. A. Korepanova. “Avtomatičeskij morfologičeskij analizator drevnerusskogo jazyka: lingvističeskie i tehnologičeskie rešenija.” V *10-ja jubilejnaja meždunarodnaja konferencija «EVA 2007 Moskva»*. Moskva, 2007.

Duhanina, Aleksandra V. *Morfologičeskie normy v sočinenijah Epifanija Premudrogo (sistema glagola)*. Avtoreferat dissertacii. Moskva, 2008.]

About the authors

Prof. Dr. Juliane Besters-Dilger is the former Head of the Department of Slavonic Studies at the University of Freiburg, Germany. Among others, her research interests concern editing Old/Middle Russian and Church Slavonic texts and glossaries, e.g. the “Commented Acts of the Apostles” (text, commentary, index of wordforms), extracted from the Great Menaion Reader of Macarius, Metropolitan of Moscow. E-mail: juliane.besters-dilger@slavistik.uni-freiburg.de

Prof. Dr. Achim Rabus is the current Head of the Department of Slavonic Studies at the University of Freiburg, Germany. Rabus defended his PhD thesis on the language of East Slavic spiritual songs in 2008 and his Habilitationsschrift on Slavic language contact in 2014. Since 2009, Rabus has been a member of the Special Commission on the Computer-Supported Processing of Mediæval Slavonic Manuscripts and Early Printed Books to the International Committee of Slavists, and since 2018, the President of the Commission. His current research focuses on Slavic social dialectology, Handwritten Text Recognition, corpus and (digital) historical linguistics. E-mail: achim.rabus@slavistik.uni-freiburg.de