BULGARIAN ACADEMY OF SCIENCES
INSTITUTE OF LITERATURE

SCRIPTA & *e*-SCRIPTA

БОЯН
ПЕНЕВ
ИЗДАТЕЛСКИ ЦЕНТЪР

This issue of the journal is published with the financial support of National Science Fund, Bulgarian Ministry of Education and Science, Grant ДНП 06/3, 15.12.2017

Scripta & *e*-Scripta

The Journal of Interdisciplinary Mediaeval Studies

Volume 18

Sofia
2018

SCRIPTA
& ⓔ- SCRIPTA

18 / 2018

Editorial Board

Editorial Address

Institute for Literature
Bulgarian Academy of Sciences
52, Shipchenski prohod Str.
1113 Sofia, Bulgaria

Executive editors of this issue:
Anissava Miltenova, Margaret Dimitrova and Elissaveta Moussakova

ISSN 1312-238X

ISSN 2603-3364 (e-publication)
“Boyan Penev” Publishing Center
Institute of Literature, BAS
Grafic design and pre-press: ***Daniela Vassileva***

Scripta & e-Scripta

Volume 18
Institute of Literature, Bulgarian Academy of Sciences
Sofia 2018

Table of Contents

e-Scripta

XVI International Congress of Slavists, Belgrade, 2018

Scripta

Conference: Mediaeval Slavic Literatures in the Digital Age: Taxonomy of Terminology and Real Text in Manuscripts

Personalia

Scripta & e-Scripta

Volume 18

Институт за литература, Българска академия на науките
София 2018

Съдържание

e-Scripta

Шестнадесети славистичен конгрес, Белград, 2018

Scripta

Конференция: Средновековните славянски литератури в дигиталната ера: таксономия на терминологията и реален текст в ръкописите

In memoriam

SCRIPTA
& e- SCRIPTA

e-SCRIPTA

18 / 2018

XVI International Congress of Slavists, Belgrade, 2018

New Developments in Tagging Pre-modern Orthodox Slavic Texts[1]

Yves Scherrer, Susanne Mocken, Achim Rabus

Abstract: Pre-modern Orthodox Slavic texts pose certain difficulties when it comes to part-of-speech and full morphological tagging. Orthographic and morphological heterogeneity makes it hard to apply resources that rely on normalized data, which is why previous attempts to train part-of-speech (POS) taggers for pre-modern Slavic often apply normalization routines. In the current paper, we further explore the normalization path; at the same time, we use the statistical CRF-tagger MarMoT and a newly developed neural network tagger that cope better with variation than previously applied rule-based or statistical taggers. Furthermore, we conduct transfer experiments to apply Modern Russian resources to pre-modern data. Our experiments show that while transfer experiments could not improve tagging performance significantly, state-of-the-art taggers reach between 90% and more than 95% tagging accuracy and thus approach the tagging accuracy of modern standard languages with rich morphology. Remarkably, these results are achieved without the need for normalization, which makes our research of practical relevance to the Paleoslavistic community.

Key words: Church Slavonic, natural language processing, part-of-speech tagging, Old Russian, neural networks

[1] Part of this work has been funded by the Innovations fonds Forschung of Freiburg University (Project “Covacs - Infrastructure: Corpus-assisted analysis of morphosyntactic variability in varieties of Church Slavonic”).

Introduction

In recent years, numerous attempts have been made to improve part-of-speech (POS) and full morphological tagging[2] of pre-modern Orthodox Slavic[3] texts. In addition to the difficulties that pre-modern Slavic texts share with other highly-inflected languages – a large tagset, numerous redundant inflectional desinences, etc. – there are two main issues that significantly complicate the task. First, despite the attempts already made, natural language processing (NLP) resources for texts written in Church Slavonic and other pre-modern Slavic varieties still leave something to be desired. Second, and more importantly, the manifold heterogeneity of pre-modern Slavic data – due to the lack of a codified norm primarily on an orthographic level, but also on a morphological level, often because of diatopic and diachronic variation – complicates the task of part-of-speech and morphological annotation to a considerable extent.

Earlier attempts to create NLP resources for pre-modern Orthodox Slavic texts follow different approaches: Some researchers develop and apply sophisticated rules for morphological analysis (Baranov et al. 2007,[4] http://bases.ruslang.ru/ by the team of the Institute for the Russian Language of the Russian Academy of Science, led by A.M. Moldovan), while others use a pipeline that includes tagging projection from Modern Russian (Meyer 2011). One of the most advanced experiments so far is based on a combination of a statistical tagger (TnT, Brants 2000) trained on a sufficient amount of normalized data together with a rule-based tagger and several pre-processing steps (Berdičevskis et al. 2016, henceforth BEG16). It yielded 92.7% correct POS tags and 81.5% correct morphology tags. Although these results are quite impressive, we still see room for improvement with new statistical and neural network taggers and normalization routines.

Our experiments follow the general setup proposed by BEG16. In particular, we use the TOROT treebank as a basis for training and testing different statistical taggers. However, we take advantage of several recent developments in corpus creation, normalization, and tagging algorithms in our experiments:

[2] We refer to ‘POS-tagging’ when aiming to predict the POS tag only, and to ‘full morphosyntactic tagging’ when aiming both to predict the POS tag and the morphosyntactic features.

[3] We use this cover term for not entirely vernacular pre-modern written texts of *Slavia Ortodossa*, including those written in Old Church Slavonic (OCS) and later recensions of Church Slavonic. Hybrid East Slavic texts traditionally labelled Old or Middle Russian, such as the Chronicles or *Domostroj*, are covered by this term as well.

[4] A web application using this morphological analyzer can be found here: http://manuscripts.ru/mns/slov.poisk?p_lang=EN

- The TOROT resource has been expanded since the reported experiments. This allows us to increase the size of the training data and the variety of test data. We also include the Old Church Slavonic dataset from the PROIEL resource in our experiments.
- In addition to the TnT tagger already used by BEG16, we experiment with the MarMoT tagger and with a tagger based on deep neural networks.
- The PROIEL corpus has been made available on the Universal Dependencies platform, after its automatic conversion to the universal part-of-speech tag, morphology, and dependency annotation formats. We have converted the TOROT data in the same way; this conversion is particularly important if cross-linguistic comparison or annotation projection is envisaged.
- While BEG16 report simple accuracy measures for morphological tagging as well as Hamming distance, we introduce a measure based on micro-averaged F1-scores (see below). This measure is more fine-grained than simple accuracy, and unlike Hamming distance, it is able to summarize tagging performance in a single value.
- The tagging experiments of BEG16 rely on a normalization routine that simplifies the original orthography by lowercasing the corpus and replacing diacritics, ligatures, and character varieties. We rely on an updated normalization routine, but also experiment with unnormalized data, a task that is of great practical relevance in Digital Paleoslavistics. Depending on the tagging algorithm, both variations yield improvements in tagging performance.
- We investigate several ways of including external resources in the tagging process, including the use of the PLDR corpus, a parallel Old Russian - Modern Russian text without annotation, and the SynTagRus treebank of Modern Russian in the Universal Dependency format. However, our experiments using these external resources do not substantially improve the tagging accuracies.

In the following sections, we describe the data, toolkits, and experiments in more detail.

Data

The main source for our experiments is the *Tromsø Old Russian and OCS Treebank*, abbreviated as TOROT (Eckhoff and Berdičevskis 2015). We use the version released on 16 June 2016, containing 24 texts with a total of approximately 242,500 word tokens.[5] The corpus consists of annotated Old Church Slavonic,

[5] https://github.com/torottreebank/treebank-releases/releases/tag/20160616

Old Russian, and Middle Russian texts from the 11th to the 16th centuries. The quantitatively most important texts in our training data are the *Codex Suprasliensis*,[6] the *Primary Chronicle* as represented in the *Codex Laurentianus*,[7] the *Domostroj* according to the *Konšinskij spisok*,[8] and excerpts of the Old Russian part of *Uspenskij Sbornik*, namely the originally East Slavic *Tale of Boris and Gleb*, and the *Life of Feodosij Pečerskij*. Note that the experiments reported by BEG16 rely on an earlier version of the corpus with approximately 175,000 word tokens.

We set aside three text chunks as test sets: the first section[9] of *Life of Sergij of Radonež* (this corresponds to the test set of BEG16 and will be used in Test 1), the first 12 sections of *Domostroj* (Test 2), and the first 68 sections of *The Primary Chronicle, Codex Laurentianus* (Test 3). These three test sets contain 1,707, 2,211, and 4,316 tokens, respectively, and represent different genres of pre-modern East Slavic literacy, as well as different times of origin. The remainder of the TOROT treebank (i.e., including the remaining sections of the three texts mentioned above) is used for training; it amounts to 234,336 tokens.

An additional data source is the Old Church Slavonic part of the *Pragmatic Resources in Old Indo-European Languages Treebank*, abbreviated as PROIEL (Haug and Jøhndal 2008, Eckhoff et al. 2018),[10] consisting of the Old Church Slavonic *Codex Marianus* Gospel manuscript (Test 4). We use the version published on the Universal Dependencies (henceforth UD) repository,[11] which provides a predefined split into training, development, and test data. The training data contains 49,862 tokens, while the test data contains 13,040 tokens.[12] We currently do not use the development data.

For our transfer learning experiments, we use two additional resources. The first one is the SynTagRus treebank of Modern Russian (Djačenko et al. 2015), in its version published on the UD repository.[13] We only use the training data in our experiments (719,535 tokens). The second resource is a parallel Old and Middle Russian - Modern Russian corpus without any morphosyntactic annotation, namely

[6] http://suprasliensis.obdurodon.org/ Editors of the digital version: Anisava Miltenova, David Birnbaum.

[7] http://pvl.obdurodon.org/ Editors of the digital version: Donald Ostrowski, David Birnbaum, Horace Lunt.

[8] http://lib.pushkinskijdom.ru/Default.aspx?tabid=5145 Editor of the digital version: Mirjam Zumstein.

[9] By "section" we understand a piece of text enclosed in a <div> tag in the XML data files.

[10] http://foni.uio.no:3000/users/sign_in

[11] https://github.com/UniversalDependencies/UD_Old_Church_Slavonic/releases/tag/r2.1

[12] We removed all verses of John 18 from the test file, as these verses are present in a similar version in the *Codex Zographensis*, which is part of the TOROT training data.

[13] https://github.com/UniversalDependencies/UD_Russian-SynTagRus/releases/tag/r2.0

the parallel data from the series *Pamjatniki literatury Drevnej Rusi*. This corpus contains 1.7 million tokens on each side and actually has some overlapping text with TOROT (though normalized in a different way). The corpus was manually aligned on the paragraph level, and paragraphs were automatically sentence-aligned and word-aligned, with the Efmaral word alignment tool (Östling and Tiedemann 2016) used for the latter.

Preprocessing

We apply two types of preprocessing to the corpora: (1) normalization of the word tokens, and (2) harmonization of the morphosyntactic annotations.

BEG16 (102) describe a normalization step that is applied to the TOROT word tokens before training and testing:

"The normalisation consists in considerable orthographical simplification. All diacritics are stripped off, all capital letters are replaced with lower-case letters, all ligatures are resolved (e.g., *ѿ* to *от*), all variant representation of single sounds are reduced to one (all *o* variants are reduced to *o* and all *i* variants are reduced to *и*, for instance). The juses are simplified to *я* and *у* (*ю*), and the jat to *е*."

Their reported results are based on taggers trained and tested on these normalized data only. Close inspection of the normalization script, which the authors have kindly shared with us,[14] showed that some diacritics, upper case letters, and Latin and Greek symbols were not correctly normalized, and that some normalization rules erroneously introduced Latin characters instead of Cyrillic characters of the same shape. We updated the normalization script accordingly and also applied it to the PROIEL and PLDR datasets. Table 1 shows some examples of normalizations.

Table 1. *Examples of word forms together with their conversion according to the normalization script used by BEG16 (norm), and to our updated normalization script (norm2).*

Unnormalized (orig)	**Normalized (norm)**	**Improved (norm2)**
ѿц҃а	отца	отца
єпиданїꙗ	єпифания	єпифания
б҄зѣ	б҄зє	бзє

[14] The supplementary material (http://hdl.handle.net/10037.1/10303) contains the normalized data, but not the scripts used to obtain them. We would like to thank Hanne Eckhoff for her support and for kindly providing us with the script. The usual disclaimers apply.

The TOROT corpus is based on the same morphosyntactic annotation guidelines and tagset as the PROIEL corpus, but this tagset has not been used in other projects, making cross-linguistic and cross-corpus studies difficult.[15] In recent years, however, the Universal Dependencies (UD) initiative has gained a lot of traction; it defines, among other things, a universal set of part-of-speech tags and morphosyntactic features, together with annotation and conversion guidelines for various languages and language families (Petrov et al. 2012, Zeman 2008, Zeman 2015).[16] As the PROIEL corpora have already been converted automatically to the UD format, the same conversion script[17] can be applied to the TOROT data as well. Table 2 shows an example of the same token annotated according to the two formats. In both annotation schemes, a token is annotated with a part-of-speech tag (the main word category) and with a set of morphosyntactic feature-value pairs. The example shows that the "interrogative" feature is encoded in the POS tag in the original scheme (through the *i* character), but in the morphosyntactic description in the UD scheme (through the *PronType=Int* feature).

Table 2. *Example of annotations for the word* коимъ *in the original PROIEL format and in the UD format.*

	Token	**POS tag**	**Morphosyntactic features**
Original	коимъ	Pi	NUMBs|GENDm|CASEi
UD	коимъ	PRON	Case=Ins|Gender=Masc|Number=Sing|PronType=Int

One nice feature of the UD initiative is that corpora from different providers, and even different languages, can be combined to create a single tagger, as the labels to be predicted are the same (e.g. Scherrer and Rabus 2017, Cotterell and Heigold 2017). In our case, we are interested in enhancing the pre-modern Slavic tagger by transferring information from a Modern Russian corpus. However, although the SynTagRus corpus and the PROIEL/TOROT corpora are annotated using the UD conventions, they diverge in several respects. Some divergences are due to the fact that they represent different language varieties in which different morphosyntactic features are relevant. But other divergences are purely due to different interpretations of the annotation and conversion guidelines. For example, PROIEL/TOROT annotate pronouns with their type (interrogative, personal, relative, etc.), whereas SynTagRus does not. On the other hand, SynTagRus distinguishes between inanimate and animate nominal forms, whereas PROIEL/TOROT do not. The negation particle

[15] As an illustration of this problem, see for example the tag harmonization paragraphs in BEG16.

[16] http://universaldependencies.org/u/feat/index.html

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

нє is considered an adverb by PROIEL/TOROT but a particle by SynTagRus. We harmonized these divergences in such a way that only the SynTagRus annotations are modified, and only where a deterministic transformation can be applied. Concretely, this means that we remove the animacy feature from SynTagRus and reannotate the negation particle as an adverb, but that we do not change anything regarding the pronoun types (as this is non-deterministic and would require manual work).

Taggers

Tagging algorithms can be used to perform either POS tagging or full morphosyntactic tagging. We use the TnT tagger (Brants 2000) as a baseline, following BEG16. This tagger has a rather simple architecture based on second-order Markov models and relies on suffix analysis to predict tags of unknown words. The TnT tagger does not provide a particular mode for full morphosyntactic tagging; instead, the POS tags are concatenated with the morphosyntactic tags before training and testing. Despite its simplicity, the TnT tagger has proven successful in various settings. In recent years, however, other more powerful tagger architectures have been proposed.

One of these tagger architectures is implemented by MarMoT (Müller et al. 2013). MarMoT uses a higher-order conditional random field (CRF) and is specially adapted to problems with large tagsets, such as full morphological tagging. Its good performance is achieved through efficient pruning techniques and coarse-to-fine tagging, i.e. it starts by predicting (coarse) part-of-speech tags, and predicts the (fine) morphosyntactic features only in a second step. Like all CRF-based models, MarMoT's predictions rely on a set of input features that are extracted from each word at training and test time. The default feature set has yielded good results in our experiments. Moreover, as shown elsewhere (Scherrer and Rabus 2017), MarMoT copes quite well with orthographic heterogeneity.

Recently, deep neural networks have been proposed for a variety of natural language processing tasks, including tagging (Collobert et al. 2011). Further refinements regarding neural networks for tagging and relevant for our own experiments were made in the following contributions:

- Wang et al. (2015) introduced recurrent neural networks, i.e. a way of forcing the tagger to make a globally optimal decision for the whole sentence by making it aware of its decisions on the previous words, similar to HMMs or CRFs. This particular form of recurrent neural network is called LSTM (long-short-term memory), or bi-LSTM for its increasingly popular bidirectional variant.
- Ling et al. (2015) proposed using character representations of the words as input features. Whereas earlier work relied on so-called word embeddings (i.e. vectors encoding the distributional similarity of words in large corpora), character

representations can be reliably obtained from smaller corpora, while nevertheless containing valuable information about morphology, which is crucial for tagging.

- Heigold et al. (2017) showed that good morphological tagging can be achieved without word embeddings, by using character representations alone.[18]
- Pinter et al. (2017) used a distinct output layer for every morphosyntactic feature, whereas previous taggers generate the whole morphosyntactic analysis at once and thus had trouble learning rare feature combinations reliably.

The architecture of a neural network tagger is defined by a large number of hyperparameters, which would take up too much space to describe in detail. We refer the interested reader to the references given above, as well as to Cotterell and Heigold (2017), whose architecture most closely resembles the one we have found to be most powerful for our task.[19] Our tagger is implemented with the DyNet toolkit (Neubig et al. 2017), based on an earlier implementation by Pinter et al. (2017).[20] In the following experiments, we call the neural network tagger CLSTM.

Evaluation

The evaluation of a part-of-speech tagger is quite straightforward: for each word, one single part-of-speech is predicted, and this prediction is either correct or wrong.[21] The value usually reported is thus part-of-speech accuracy, i.e. the proportion of correct predictions among all predictions.

However, things are less clear for full morphological tagging. Let us assume the correct gold tag[22] (1) and potential predictions (2), (3), and (4):

[18] There is still an ongoing debate on whether character representations are sufficient for tagging or not: while Heigold et al. (2017) argue in favor of purely character-level models, Horsmann and Zesch (2017) find improvements for a large number of languages with added word embeddings.

[19] In particular, we use the following settings: character representations are created using two bi-LSTM layers with a character input vector of 128 dimensions and hidden layers of 256 dimensions; word embeddings are not used; the tagger itself consists in two bi-LSTM layers with hidden layers of 256 dimensions; we use distinct output layers for each attribute. The models are trained for 40 epochs (i.e., 40 complete passes through the training data) with the MomentumSGD algorithm, dropout of 0.02, and learning rate decay of 0.1.

[20] https://github.com/yvesscherrer/lstmtagger

[21] It could be argued that predicting, for example, a proper noun instead of a common noun is somewhat less serious than predicting e.g. an adverb instead of a common noun, but such an evaluation would require a distance table between every pair of labels, which is difficult to compile.

[22] In NLP, the manual annotations that are assumed to be correct and with which the output of automatic tools is compared are called “gold annotations” or “gold standard”.

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

(2) Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin |Voice=Act

(3) Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin

(4) Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Gender=Masc|Tense=Past|VerbForm=Fin|Voice=Act

The prediction (2) deviates from (1) only by 1 out of 7 feature values. Calculating simple accuracy would therefore be rather harsh, as any two non-matching strings will count as a wrong prediction. A finer-grained measure is therefore needed. One possibility is to compute accuracy feature-wise, according to which the prediction (2) would count as 6/7=85.7% correct. However, feature-wise accuracy is difficult to compute in cases of missing (3) or excessive (4) features. Following Pinter et al. (2017), we report micro-averaged attribute F1 scores. This computation works in the following way:

1. Count the number of attributes (let us call a feature-value pair an 'attribute') in the gold tag and in the predicted tag, and the number of common attributes ("correct attributes").
2. Compute the precision (number of correct attributes / number of gold attributes) and the recall (number of correct attributes / number of predicted attributes).
3. Compute the F1-score, defined as the harmonic mean of precision and recall (2 * recall * precision / (recall + precision)).
4. Average the F1-scores over all examples of the corpus.

With respect to the examples above, this computation yields the following values:

(2) Precision: 6/7=0.857, Recall: 6/7=0.857, F1: 0.857

(3) Precision: 6/6=1.0, Recall: 6/7=0.857, F1: 0.923

(4) Precision: 7/8=0.875, Recall: 7/7=1.0, F1: 0.933

These examples show that micro-F1 is equivalent to attribute-wise accuracy if the number of predicted attributes equals the number of gold attributes. It also shows that excessively predicted attributes are penalized less than missing attributes, which in turn are penalized less than wrongly predicted attributes. When comparing with earlier work, we also report simple accuracy.

Another important figure we report is the out-of-vocabulary (OOV) rate. It shows what percentage of tokens of the test set has not been seen during training. The underlying idea is that OOV words are harder to tag than previously seen words and that reducing the OOV rate should result in improved tagging. This idea motivates the use of normalization procedures in particular in order to reduce the OOV rate by reducing orthographic variability. The experiments will show, however, that this hypothesis does not necessarily hold any more with new tagging models.

Experiments

In the first set of experiments, we train and test our models on the TOROT resource only, keeping the original annotation scheme. We report the results for the three tagging algorithms and for the three types of normalization.

Table 3. *Tagging experiments with the TOROT training data in the original annotation format. The best figure of each column is displayed in bold.*

Tagger	Norm.	Test 1: Sergrad			Test 2: Domo		Test 3: Lav	
		POS Acc	Morph Acc	Morph MicroF1	POS Acc	Morph MicroF1	POS Acc	Morph MicroF1
TnT (BEG16)	norm	89.5%	81.5%					
TnT	orig	90.57%	83.54%	89.43%	94.26%	90.73%	89.50%	88.96%
	norm	92.56%	86.06%	91.33%	95.12%	92.04%	90.43%	89.30%
	norm2	93.03%	86.35%	91.83%	95.02%	92.11%	90.43%	89.31%
MarMoT	orig	93.44%	86.94%	91.94%	95.34%	92.93%	91.29%	91.48%
	norm	94.67%	89.05%	93.38%	96.47%	94.05%	**91.77%**	91.86%
	norm2	95.08%	89.10%	93.45%	96.47%	93.80%	91.61%	91.96%
CLSTM	orig	**95.78%**	**90.69%**	94.86%	96.70%	93.68%	91.03%	**92.55%**
	norm	95.08%	89.46%	94.39%	96.29%	93.83%	91.66%	91.79%
	norm2	95.37%	90.39%	**95.11%**	**96.83%**	**94.17%**	90.96%	91.78%

Table 3 shows several results. First, for all except one number, the CLSTM tagger performs best, although MarMoT does not trail far behind. In contrast, the results of the TnT tagger are several percentage points lower. Second, the impact of normalization depends on the tagger used: whereas normalization always helps for TnT, its impact is weaker for MarMoT, and CLSTM often even performs best with unnormalized data. This suggests that CLSTM, and to a certain extent also MarMoT, is better at generalizing from non-standard spellings than TnT. The difference between norm and norm2 is rather small, but generally in favor of norm2. Third, if we compare the different test corpora, it appears that Test 1 and Test 2 are easier to annotate than Test 3. This can be explained by the fact that in the *Primary Chronicle* there are numerous proper nouns, often in enumerations such as:

> Сущимъ же ко востокомъ имать киликию памъфилию писидию мосию лукаѡнию фругию камалию ликию карию лудью масию другую троаду салиду вифунию старую фругию.
>
> Towards the Orient, there are Cilicia, Pamphylia, Pisidia, Mysia, Lycaonia, Phrygia, Camalia, Lycia, Caria, Lydia, Moesia Secunda, Troas, Aeolia, Bithynia, and ancient Phrygia.

The taggers often fail to recognize the part of speech PROPN (proper noun) correctly and so replace it with NOUN or, in some cases, another tag. However, since PROPN is a subtype of NOUN, morphological interpretation is often correct nevertheless, which explains the – at first sight counter-intuitive – result that, in Test 3, morphology micro-F1 is consistently higher than POS accuracy.

Finally, we can compare our results to those reported by BEG16. In a comparable setting (TnT + norm), we were able to increase POS accuracy by 3% and morphology accuracy by 4.5%, thanks to the larger training corpus available. Also, we could easily surpass the best reported POS accuracy in BEG16 (92.7%, achieved through combination of their TnT tagger with a rule-based morphological guesser). Note that morphology accuracy is about 4-5% lower than morphology micro-F1, because of its "harshness" discussed above.

In a second set of experiments, we turn to the TOROT corpus converted to Universal Dependency format. We only report results for the three most successful settings of Table 3, i.e. MarMoT/norm2, CLSTM/orig, and CLSTM/norm2. In addition, we include the Old Church Slavonic PROIEL training data in our experiments (either on its own or concatenated with TOROT), and add the PROIEL test set as Test 4.

Table 4. *Tagging experiments with different combinations of TOROT and PROIEL training data in the UD annotation format. Results in boldface represent per-column maxima, results on grey background represent the best results with TOROT+PROIEL settings.*

Tagger	Training	Test 1: Sergrad		Test 2: Domo		Test 3: Lav		Test 4: Marian	
		POS	Mor	POS	Mor	POS	Mor	POS	Mor
MarMoT norm2	TOROT	94.55	93.12	**96.34**	93.24	91.57	91.48	92.95	90.22
	PROIEL	76.16	71.60	78.15	69.08	67.17	66.01	95.12	93.66
	TOROT+PROIEL	94.20	93.16	96.20	93.20	91.45	91.41	95.95	94.47
CLSTM orig	TOROT	95.02	93.88	95.79	**93.60**	**91.89**	**91.94**	94.07	91.22
	PROIEL	75.28	68.98	78.02	65.92	72.64	63.64	95.50	94.27
	TOROT+PROIEL	94.84	94.18	95.84	92.22	90.64	91.54	**96.16**	**95.01**
CLSTM norm2	TOROT	**95.25**	**94.29**	96.25	93.57	90.11	91.61	93.60	91.16
	PROIEL	83.95	76.49	86.07	75.03	74.24	70.01	95.44	93.95
	TOROT+PROIEL	94.84	94.14	96.07	93.06	90.43	90.89	95.80	94.10

Generally, the results using the UD annotation scheme are a bit lower than with the original annotation scheme. This might be the result of some minor inconsistencies introduced by the automatic conversion process.

In terms of training/test data combination, one could expect that the test sets from TOROT (Tests 1-3) would be best tagged with taggers trained on TOROT data, and the PROIEL test set with taggers trained on PROIEL data. This expectation is met, although there are some striking imbalances: while annotating Test 4 (test data from the OCS *Codex Marianus*) with the "wrong" tagger only decreases results by at most 3%, using the "wrong" tagger on Test 1-3 decreases results by more than 20%. This is in line with the OOV rates reported in Table 5 below, suggesting that the TOROT training resource is more diverse and more resilient to test data from other sources. Indeed, the TOROT corpus encompasses 24 different texts with different linguistic characteristics, whereas the whole PROIEL corpus comes from the same Old Church Slavonic text source; also, TOROT is nearly 5 times bigger than PROIEL. From a linguistic viewpoint, this result makes sense as well, since TOROT encompasses Old Church Slavonic texts (namely *Codex Zographensis* and *Codex Suprasliensis*), whereas PROIEL does not contain any Old Russian or Middle Russian training data.

When combining the two training corpora, tagging results typically improve on Test 4, but decrease slightly on Test 1-3. We intend to study in further detail the issue of finding the ideal trade-off between size and linguistic homogeneity of training data in future experiments.

Regarding the three tagger settings, the results are not very clear. CLSTM[23] performs best on Tests 1, 3, and 4, but MarMoT is better on Test 2. Normalization helps for Test 1, but not for Tests 3 and 4 – possibly a consequence of the (implicit or explicit) tuning of the normalization routine to the initial test set Test 1 – whereas the impact of normalization on Test 2 is quite unclear. It is interesting to note that taggers based on neural networks are powerful enough to obviate the need for normalization, so that normalization even harms tagging performance in some cases. This means that the use of state-of-the-art taggers such as CLSTM and, to a lesser extent, MarMoT can solve the most severe issue when it comes to tagging pre-modern Slavic texts, i.e. orthographic heterogeneity.

In order for the Digital Philologist to get a picture of the tagging performance, we present randomly selected sentences from the *Primary Chronicle* that were tagged with CLSTM. Incorrectly tagged features are marked in italics:

и	CCONJ
брака	NOUN *Case=Nom*|*Gender=Fem*|Number=Sing OOV
оу	ADP
нихъ	PRON Case=Gen|Gender=Masc|Number=Plur|Person=3|PronType=Prs
не	ADV Polarity=Neg
бываше	VERB Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act

"… and there was no marriage among them …" (Cross and Sherbowitz-Wetzor 1953: 56)

Here, CLSTM misinterpreted брака as a feminine noun and did not recognize the grammatically homonymous genitive of negation of the masculine noun бракъ 'marriage'.

аще	SCONJ
кто	ADJ Case=Nom|Number=Sing

[23] It should be noted, however, that the CLSTM results are dependent on the particular hyperparameter settings chosen, and that careful tuning of these parameters could further improve tagging accuracy.

оумраше	VERB Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act OOV
твораху	VERB Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Past|Verb Form=Fin|Voice=Act
трызно	*ADV* OOV
надъ	ADP
нимъ	PRON Case=Ins|Gender=Masc|Number=Sing|Person=3|PronType=Prs
и	CCONJ
по	ADP
семь	ADJ Case=Loc|Gender=Neut|Number=Sing
твораху	VERB Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin|Voice=Act
кладу	NOUN *Case=Dat*|*Gender=Masc*|Number=Sing OOV
велику	ADJ *Case=Dat*|Degree=Pos|*Gender=Masc*|Number=Sing|Strength=Strong
и	CCONJ
възложахуть	VERB Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin|Voice=Act OOV
и	PRON Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs
на	ADP
кладу	NOUN *Case=Loc*|*Gender=Masc*|Number=Sing OOV

“Whenever someone died, they held a feast over him, and afterwards they created a large pyre and laid him onto the pyre.”

CLSTM correctly discriminates between the conjunction и ‘and’ and the pronoun и ‘him’ and is able to recognize both the etymological imperfect form твораху ‘they created’ and the East Slavic innovation възложахуть ‘they laid’ correctly. It thus demonstrates an overall good tagging performance that is not restricted to words learned during the training session (cf. the correctly tagged OOV-tokens оумраше ‘he died’ and възложахуть ‘they laid’). However, CLSTM fails to recognize the neuter noun трызно ‘feast’[24] and the gender (and, as a result, case) of the phrase кладу велику ‘great pyre’.

[24] In the gold standard, the token is tagged as neuter in the accusative. However, the variants read трызну, and Sreznevskij (s.v. тризна) cites the *Chronicle* as Аще кто умраше, твораху тръізно(у), i.e. as a feminine noun in the accusative as well.

<table>
<tr><td>и</td><td>CCONJ</td></tr>
<tr><td>баста</td><td>VERB Aspect=Imp|Mood=Ind|Number=Dual|Person=3|Tense=Past|VerbForm=Fin|Voice=Act</td></tr>
<tr><td>оу</td><td>ADP</td></tr>
<tr><td>нєго</td><td>PRON Case=Gen|Gender=Masc|Number=Sing|Person=3|PronType=Prs</td></tr>
<tr><td>в҃</td><td>NUM</td></tr>
<tr><td>мужа</td><td>NOUN Case=Nom|Gender=Masc|Number=Dual</td></tr>
<tr><td>нє</td><td>ADV Polarity=Neg</td></tr>
<tr><td>плємєни</td><td>NOUN <i>Case=Nom</i>|Gender=Neut|Number=Sing OOV</td></tr>
<tr><td>єго</td><td>PRON Case=Gen|Gender=Masc|Number=Sing|Person=3|PronType=Prs</td></tr>
<tr><td>ни[25]</td><td>CCONJ</td></tr>
<tr><td>боярина</td><td>NOUN Case=Nom|Gender=Masc|<i>Number=Sing</i> OOV</td></tr>
</table>

“And there were two men with him, neither of his kin nor boyars”

Here, the OOV tokens are problematic as well: the genitive of the n-stem noun плємєни ‘kin’ – irregular from a synchronic point of view – is, for some reason, rendered as a nominative. Moreover, while the dual of the learned token мужа is recognized correctly, CLSTM fails to recognize the analogous dual form боярина ‘two boyars’.

Table 5. *OOV rates for the tagging experiments reported in Tables 3 and 4.*

Training	Normal.	Test 1	Test 2	Test 3	Test 4
TOROT	orig	27.48%	20.81%	20.99%	29.09%
	norm	20.80%	15.92%	18.54%	16.50%
	norm2	18.04%	15.92%	18.49%	16.49%
PROIEL	orig	57.00%	60.65%	61.31%	14.28%
	norm2	46.16%	48.21%	52.34%	13.03%
TOROT+ PROIEL	orig	27.24%	20.67%	20.74%	11.83%
	norm2	17.87%	15.60%	18.16%	8.48%

[25] This is the reading in *Lavrent’evskaja Letopis’*; other witnesses read но ‘but’.

Table 5 shows the out-of-vocabulary rates for the experiments reported in Tables 3 and 4. Normalization always reduces the out-of-vocabulary rate, but the differences between norm and norm2 are marginal. At least for TnT and MarMoT, there is a good (negative) correlation between OOV rates and tagging performance.[26] In contrast, for a character-based neural network tagger like CLSTM, there is no real difference between OOV tokens and in-vocabulary tokens: the internal representations are constructed in the same way for all tokens, character by character, and OOV tokens just happen to combine the characters in a way that has not been seen during training time. As a result, there is no clear correlation between tagging OOV rates and tagging performance. Nevertheless, as shown in the qualitative philological analysis, even the neural tagger CLSTM would benefit from additional resources that reduce OOV rates. What exactly such resources should look like remains to be investigated: there is no direct path in CLSTM from input word representations to output tags that could be complemented with "shortcuts" from a morphological dictionary.

Another interesting line of future research would concern simultaneous tagging and normalization using the same neural network model.

Transfer learning experiments

In this section, we present four experiments that use the additional resources PLDR and SynTagRus in different combinations. The general idea of these experiments is to take advantage of the additional Old Russian data of PLDR, of the correspondence between Old and Middle Russian and Modern Russian encoded in the PLDR parallel corpus, and of the token-tag assignments available in the Modern Russian SynTagRus corpus. The experiments below use the MarMoT tagger, as it is faster to train and provides more options to include additional data. Also, if not stated otherwise, we use the norm2 normalization.

Modernizing the training and test data

An often-used strategy for tagging historical corpora without training data is to modernize the historical text, tag it with a tagger trained on modern-language data, and project the tags back to the historical original (Scherrer and Erjavec 2016, Tjong

[26] The Pearson correlation between OOV rate and tagging performance amounts to -0.51 and -0.60 for TnT (POS-tagging and morphosyntactic tagging respectively) and to -0.34 and -0.61 for MarMoT, but to +0.04 and +0.22 for CLSTM.

Kim Sang et al. 2017). Various methods have been proposed for modernization, among them character-level statistical machine translation (CSMT). We replicate this setting here, using the PLDR corpus to train the modernization tool and training a MarMoT tagger on the Modern Russian SynTagRus corpus, and we then use this tagger to annotate the modernized test sets.

Recall that the PLDR corpus contains normalized Old and Middle Russian text on one side, and Modern Russian text on the other side, and that both sides are aligned at the sentence level. This dataset is used to train a modernization system according to the CSMT paradigm, using the same settings as for normalizing spellings of modern dialects (Scherrer and Ljubešić 2016), modernizing historical data (Ljubešić et al. 2016), and for normalizing user-generated content such as tweets (Ljubešić et al. 2016).

In contrast to other language settings, we are in the comfortable situation of having pre-modern Slavic training data, namely the TOROT and PROIEL training sets used in the experiments above. It would be unfortunate to forgo these resources. Hence, in an additional experiment, we modernize the TOROT and PROIEL training sets using the same model and concatenate this training data with the SynTagRus corpus before creating the MarMoT tagger.

Table 6. *Tagging with automatically modernized data.*

Tagger	Training	Test 1: Sergrad		Test 2: Domo		Test 3: Lav		Test 4: Marian	
		POS	Mor	POS	Mor	POS	Mor	POS	Mor
MarMoT	SynTagRus	71.33	62.18	78.98	62.59	65.15	56.88	64.12	56.95
	TOROT+ PROIEL+ SynTagRus	86.72	80.74	92.66	85.69	85.70	79.68	88.31	84.52

The results obtained with this approach are considerably lower than those obtained with the purely supervised taggers trained on TOROT and PROIEL alone. This is, at least for the first experiment, not surprising: using projected data from another language variety is always a "poor man's solution" that is usually applied in truly low-resource scenarios and will not result in better tagging performance than a regular supervised tagger. However, the failure of the second experiment of Table 6 is more surprising. Three factors may explain this result. First, the modernization is noisy and leads to some wrongly modernized, and consequently wrongly tagged, words. Second, modernization may remove some truly useful linguistic hints, making the tagging problem more difficult. Third, even if the annotations are harmonized between the different resources, the SynTagRus corpus may differ so much in text genre and vocabulary domain that it would not constitute a good training resource even if the modernization tool worked perfectly.

Creating additional synthetic training data

In the previous experiment, we used the PLDR data only to train the modernization system. Here, we intend to use this dataset to create an additional synthetic Old Russian training corpus, roughly following earlier work by Meyer (2011). In order for this data to be usable for training, it needs to be annotated with POS tags and morphological features first. This pre-annotation is done in four phases:

1. The modern side of the PLDR corpus is annotated with a MarMoT tagger trained on SynTagRus (the same as in the previous section).
2. The Old Russian side of the PLDR corpus is annotated with a MarMoT tagger trained on TOROT+PROIEL (the same as in the initial experiments above).
3. For all Old Russian tokens that are aligned with a single Modern Russian token and which are similar enough one to another (we use a relative Levenshtein distance value of 0.3 as a similarity threshold), the TOROT+PROIEL annotation is replaced by the SynTagRus annotation projected from the aligned Modern Russian token.
4. Punctuation signs and numerals are assigned the correct tags on the basis of a small dictionary (the pre-modern Slavic training corpora do not contain such symbols, which makes the initial tagging error-prone).

The table below lists some examples.

Table 7. *Examples of the pre-annotation algorithm.*

Old word	Aligned modern word	TOROT tag associated with old word	SynTagRus tag associated with modern word	Decision
отидє	миновал	VERB:Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act	VERB:Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act	Levenshtein distance is higher than 0.3, so take TOROT tag
вся	всех	DET:Case=Acc|Gender=Masc|Number=Plur	DET:Case=Loc|Number=Plur	Take SynTagRus tag
»	»	CCONJ	PROPN:Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing	PUNCT, as in dictionary

This pre-annotation yields a corpus of 1.74 million tokens, of which 46% are annotated using the transferred SynTagRus tags, 36% using the TOROT+PROIEL tags, and 18% using the punctuation dictionary. This corpus is then concatenated with the original TOROT+PROIEL corpora to train a new MarMoT tagger. The results of this new tagger are presented below.

Table 8. *Results of a tagger trained on additional synthetic training data.*

Tagger	Training	Test 1: Sergrad		Test 2: Domo		Test 3: Lav		Test 4: Marian	
		POS	Mor	POS	Mor	POS	Mor	POS	Mor
MarMoT	TOROT+ PROIEL+ PLDR	89.22	87.21	93.67	87.56	88.14	86.32	94.31	92.97

The obtained results again lie well below those obtained without the additional PLDR training data. This is, however, not surprising given the training and test configurations of our experiments. Since the test data originate from the TOROT and PROIEL collections, the best possible tagger is one that is trained on data of the same genre and origin and with the same transcription and annotation conventions. The PLDR data, originating from a different source, follows slightly different transcription guidelines. Likewise, the reservations concerning the use of the SynTagRus corpus made in the previous section still hold in this experiment. In addition, the pre-annotation of the PLDR corpus is not error-free, leading to somewhat noisy training data. Despite these negative results, we believe that there is value in this kind of tagger: because it has been trained on data from different sources, we expect it to be more robust to test data from other sources. Unfortunately, as of now, we have not been able to test this hypothesis.

Creating a morphological dictionary

The initial experiments show a clear correlation between the OOV rate and the tagging accuracy, at least for traditional taggers such as TnT or MarMoT (see Table 5). A standard technique to reduce the OOV rate is to include a morphological dictionary for OOV words, in which all possible analyses for each OOV word are given. In our case, there are also ambiguous tokens that have only been seen in one of the possible analyses during training. Such tokens do not count as OOV words, but could benefit from extended coverage of a morphological dictionary as well. We create a morphological dictionary in the following way:

1. Extract the list of all word types of the test corpora.
2. Modernize these words using the character-level machine translation system presented above.[27]
3. If the resulting modernized form occurs in the SynTagRus corpus, create a dictionary entry consisting of the original word form and the SynTagRus annotation(s).[28]
4. Annotate the test data using a MarMoT tagger with the additional dictionary[29] created in steps 1-3.

The resulting dictionary covers 5,420 out of 5,878 word types of the test corpora, with a total of 8,383 entries (i.e. each word type has 1.55 entries on average).

A different option consists in directly using the word alignment of PLDR instead of the CSMT system to find the correspondences. According to this option, the dictionary is created as follows:

1. Extract the list of all word types of the test corpora.
2. Retrieve these words in the Old Russian part of PLDR. If found, check whether it is aligned with a single Modern Russian word and whether they are similar enough (same Levenshtein threshold as above). If so, add the SynTagRus annotation(s) associated with the Modern Russian word as candidate annotation.
3. Create a dictionary entry by choosing the most frequently seen candidate annotation.
4. Annotate the test data using a MarMoT tagger with the additional dictionary created in steps 1-3.

The resulting dictionary covers 3,063 out of 5,878 word types of the test corpora, with exactly one entry per word type.

[27] In practice, we use a slightly different CSMT system here: we use a system that is trained on single words only, instead of entire sentences as above. This has shown better performance for this particular task.

[28] For each token, we generate the 100 best modernized forms with CSMT and consider the first form that occurs in the SynTagRus corpus. If none of the 100 forms occur in SynTagRus, we skip the token.

[29] MarMoT conveniently proposes the *-type-dict* option for including the dictionary.

Table 9. *Tagging results with additional morphological dictionaries.*

Tagger	Training	Test 1: Sergrad		Test 2: Domo		Test 3: Lav		Test 4: Marian	
		POS	Mor	POS	Mor	POS	Mor	POS	Mor
MarMoT	TOROT+ PROIEL+ CSMT-Dict	94.32	92.85	96.25	93.39	91.68	91.83	96.02	94.45
	TOROT+ PROIEL+ Align-Dict	95.02	93.31	96.11	93.32	91.29	91.64	96.02	94.49

Both dictionary generation methods seem to yield equivalent results. The taggers with added dictionaries produce slightly better results than the one without a dictionary, but the differences are not statistically significant.[30]

Conclusion

Our experiments convincingly show that pre-modern Slavic texts (including texts written in OCS and Old and Middle Russian) can be tagged with an accuracy between 90% and more than 95%. This is close to the state-of-the art tagging performance for modern languages with rich morphology and a high amount of training data. Obviously, for modern taggers such as MarMoT or CLSTM, the size of the training corpora TOROT and PROIEL is sufficient to approach truly high-resourced languages. Our most relevant result for practical application is that taggers such as MarMoT and especially the neural network tagger CLSTM do not need to rely on normalized data to achieve good results. Paleoslavists can thus skip the tedious, sometimes idiosyncratic and error-prone procedure of normalization when preparing tagged corpora. If normalization is required for other purposes (e.g. corpus search), one could even imagine including a normalization output layer in the neural network tagger, so that tagging and normalization are learned by the same model.

However, the different transfer learning experiments have shown that it is not easy to improve over the baseline taggers when the training data are of consequential size. We assume, nevertheless, that the inclusion of training material from other sources would create a tagger that is less sensitive to the training genre and origin, but further work will be required to test this hypothesis.

[30] Statistical differences, computed as chi-square p-values, between the baseline MarMoT tagger and CSMT-Dict tagger range from 0.29 to 0.95, and between the baseline and Align-Dict tagger from 0.56 to 0.98.

Moreover, future work will include tests with other (especially South Slavic) premodern Slavic texts. Finally, it would be of great practical value to adapt the CLSTM tagger to be used in real-world applications, which includes rules for punctuation or accents with eventual integration into web servers and graphical user interfaces.

LIST OF ABBREVIATIONS

Acc	Accuracy
BEG16	Berdičevskis et al. 2016
CLSTM	Character-based long-short-term-memory [tagger]
CRF	Conditional Random Field
CSMT	Character-level statistical machine translation
Domo	Domostroj
HMM	Hidden Markov Models
Lav	Lavrent'evskaja Letopis'
Marian	Codex Marianus
MicroF1	Micro-averaged F1-scores
NLP	Natural language processing
OCS	Old Church Slavonic
OOV	Out of vocabulary
PLDR	Pamjatniki literatury Drevnej Rusi
POS	Part-of-speech
PROIEL	Pragmatic Resources in Old Indo-European Languages Treebank
Sergrad	Life of Sergij of Radonež
TnT	Trigrams'n'Tags
TOROT	Tromsø Old Russian and OCS Treebank
UD	Universal Dependencies

REFERENCES

Baranov et al. 2007: Baranov, V., A. Mironov, A. Lapin, I. Mel'nikova, A. Sokolova and E. Korepanova. "Avtomatičeskij morfologičeskij analizator drevnerusskogo jazyka: lingvističeskie i technologičeskie rešenija." In *10-ja jubilejnaja meždunarodnaja konferencija "EVA 2007 Moskva"*. Moskva, 2007.

Berdičevskis et al 2016: Berdičevskis, A., H. M. Eckhoff and T. Gavrilova. "The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian." In *Computational Linguistics and Intellectual Technologies. Proceedings of Dialogue 16.* Moscow, 2016, 99–111.

Brants 2000: Brants, T. “TnT – A Statistical Part-of-Speech Tagger.” In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. Seattle, WA, 2000, 224–231.

Collobert et al. 2011: Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. “Natural Language Processing (Almost) from Scratch.” *Journal of Machine Learning Research* 12 (2011): 2493–2537.

Cotterell and Heigold 2017: Cotterell, R. and G. Heigold. “Cross-lingual, Character-Level Neural Morphological Tagging.” In *Proceedings of EMNLP*. Stroudsburg, PA, 2017, 748–759.

Cross and Sherbowitz-Wetzor 1953: Cross, S. H. and O. P. Sherbowitz-Wetzor (eds.). *The Russian Primary Chronicle. Laurentian Text*. Cambridge, MA, 1953.

Djačenko et al. 2015: Djačenko, P., L. Iomdin, A. Lazurskij, L. Mitjušin, O. Podlesskaja, V. Sizov, T. Frolova and L. Cinman. “Sovremennoe sostojanie gluboko annotirovannogo korpusa tekstov russkogo jazyka (SinTagRus).” In *Sbornik “Nacional’nyj korpus russkogo jazyka: 10 let proektu”*. Moskva, 2015, 272–299.

Eckhoff and Berdičevskis 2015: Eckhoff H. and A. Berdicevskis. 2015. “Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank.” *SeS* 14-15 (2015): 9–25.

Eckhoff et al. 2018: Eckhoff, H., K. Bech, G. Bouma, K. Eide, D. Haug, O. E. Haugen and M. Jøhndal. “The PROIEL treebank family: a standard for early attestations of Indo-European languages.” In *Language Resources and Evaluation* 52/1 (2018): 29–65.

Haug and Jøhndal 2008: Haug, D. T. T. and M. L. Jøhndal. “Creating a Parallel Treebank of the Old Indo-European Bible Translations.” In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data* (LaTeCH 2008), ed. by C. Sporleder and K. Ribarov. Marrakech, 2008, 27–34.

Heigold et al. 2017: Heigold, G., G. Neumann and J. van Genabith. “An extensive empirical evaluation of character-based morphological tagging for 14 languages.” In *Proceedings of EACL*. Stroudsburg, PA, 2017, 505–513.

Horsmann and Zesch 2017: Horsmann, T. and T. Zesch. “Do LSTMs really work so well for PoS tagging? A replication study.” In *Proceedings of EMNLP*. Stroudsburg, PA, 2017, 727–736.

Ljubešić et al. 2016. Ljubešić, N., K. Zupan, D. Fišer and T. Erjavec. “Normalising Slovene data: historical texts vs. user-generated content.” In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, ed. by S. Dipper (Bochumer Linguistische Arbeitsberichte, 16). Bochum, 2016, 146–155.

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

Müller et al. 2013: Müller, T., H. Schmid and H. Schütze. “Efficient Higher-Order CRFs for Morphological Tagging.” In *Proceedings of EMNLP*. Stroudsburg, PA, 2013, 322–332.

Neubig et al. 2017: Neubig G., C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui,

C. Gan, D. Garrette, Y. Ji, L. Kong, A. Kuncoro, G. Kumar, Ch. Malaviya, P. Michel, Y. Oda, M. Richardson, N. Saphra, S. Swayamdipta and P. Yin. "DyNet: The Dynamic Neural Network Toolkit." arXiv preprint <arXiv:1701.03980>.

Östling and Tiedemann 2016: Östling, R. and J. Tiedemann. "Efficient word alignment with Markov Chain Monte Carlo." *The Prague Bulletin of Mathematical Linguistics* 106 (2016): 125–146.

Petrov et al. 2012: Petrov, S., D. Das and Ryan McDonald. "A universal part-of-speech tagset." In *Proceedings of LREC.* Istanbul*,* 2012, 2089–2096.

Pinter et al. 2017: Pinter, Y., R. Guthrie and J. Eisenstein. "Mimicking word embeddings using subword RNNs." *Proceedings of EMNLP*. Stroudsburg, PA, 2017, 102–112.

Scherrer and Erjavec 2016: Scherrer, Y. and T. Erjavec. "Modernising historical Slovene words." *Natural Language Engineering* 22(6) (2016): 881–905.

Scherrer and Ljubešić 2016: Scherrer, Y. and N. Ljubešić, "Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation." In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, ed. by S. Dipper (Bochumer Linguistische Arbeitsberichte, 16). Bochum, 2016, 248-255.

Scherrer and Rabus 2017: Scherrer, Y. and A. Rabus. "Multi-source morphosyntactic tagging for spoken Rusyn." In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Stroudsburg, 2017, 84–92.

Sreznevskij 1893ff: Sreznevskij, I.: *Materialy dlja slovarja drevne-russkago jazyka po pis'mennym pamjatnikam.* T. I-III. Sanktpeterburg, 1893ff.

Tjong Kim Sang et al. 2017: Tjong Kim Sang, Erik et al. "The CLIN27 Shared Task: Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation." *Computational Linguistics in the Netherlands* 7 (2017): 53–64.

Wang et al. 2015: Wang, P., Y. Qian, F. K. Soong, L. He and H. Zhao. "Part-of-speech tagging with bidirectional long short-term memory recurrent neural network." 2015, ArXiv preprint <abs/1510.06168>.

Zeman 2008: Zeman, D. "Reusable Tagset Conversion Using Tagset Drivers." In *Proceedings of LREC*, Marrakech, 2008, 213–218.

Zeman 2015: Zeman, D. "Slavic Languages in Universal Dependencies." In *Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning. Bratislava, Slovakia*. Bratislava, 2015, 151–163.

About the authors…

Dr Yves Scherrer has been a post-doctoral researcher in Language Technology at the University of Helsinki since March 2017. Prior to that, he held post-doctoral research and teaching positions at the University of Geneva (2013-2017) and at the University Paris 7 Diderot (2012-2013). He defended his PhD thesis on the computational modelling of Swiss German dialects, with an emphasis on machine translation techniques, in 2012 at the University of Geneva. In 2009, he was granted a Swiss National Science Foundation Fellowship for Prospective Researchers for a year-long visit to Columbia University, New York. Yves Scherrer has been involved in various projects in the areas of language technology, dialectology, and corpus linguistics. His current research focuses on the automatic analysis and annotation of variational data (such as dialectal and diachronic data), machine translation, crowdsourced data collection, and the dialectometrical analysis of corpora and inquiries. He has worked with Germanic, Romance, Slavic, and Finno-Ugric languages.

Dr Susanne Mocken currently works as a research fellow in IT Services at the University of Freiburg, Germany. She holds a Master's Degree in Romance and German Philology. She finished her PhD thesis in 2013, which involved analysing the presentation of direct speech in 70 French novels using computational methods. For this work, she received the Irmgard Ulderup Award for outstanding research in Romance Studies in 2014. Since then, she has worked for several projects funded by the Federal Ministry of Education and Research, the German Research Foundation, the federal state of Baden-Württemberg, and the University of Freiburg. Her focus is currently on establishing IT structures for fellow researchers, research data management, programming, and project coordination.

Prof. Dr Achim Rabus holds the Chair of Slavic Linguistics at the University of Freiburg, Germany. From 2013 until 2016 he was employed as a Professor at the University of Jena and the Aleksander Brückner Center for Polish studies at the Universities of Halle and Jena. Rabus defended his PhD thesis on the language of East Slavic spiritual songs in 2008 and his Habilitationsschrift on Slavic language contact in 2014. In 2012 he was awarded a Feodor Lynen fellowship to conduct research at the University of California, Berkeley, sponsored by the Alexander von Humboldt Foundation. From 2011 until 2016, he was a member of the Junior Academy Program of Heidelberg Academy of Sciences and Humanities. 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 2016, of the Working Group F2 (other philologies) of CLARIN-D. He has been involved in several philological, sociolinguistics, digital humanities, and corpus linguistics projects. His current research focuses on Slavic sociolinguistics, dialectology, corpus and (digital) historical linguistics.