

# Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus<sup>1</sup>

Achim Rabus

**Abstract:** The paper discusses the automatic text recognition capabilities of neural network models specifically trained to recognize different styles of Church Slavonic handwriting within the software platform Transkribus. Computed character error rates of the models are in the range of 3 to 4 percent; real-life performance shows that specifically trained models, by and large, recognize simple (non-superscript) characters correctly most of the time. Error rate is higher with superscript letters, abbreviations, and word separation. Combined models consisting of training data from different sources are capable of transcribing different styles of Slavic handwriting with low error rate. Automatic text recognition using Transkribus and the models presented in this paper can help improve the efficiency of the process of digitizing Church Slavonic manuscripts and, thus, boost the number of digitized sources available in the future.

**Key words:** Church Slavonic, Transkribus; automatic transcription, machine learning, neural networks, artificial intelligence

## Introduction

Digital text recognition of Church Slavonic manuscripts in order to make their text digitally available is a difficult task due to several reasons: First, there is huge variation in writing styles with respect to both orthography and the rendition of glyphs. While the difference between, say, *ustav*, and *poluustav*, does not prevent a trained philologist from correctly recognizing the manuscript text, it may become an insurmountable obstacle to a piece of software. Second, many manuscripts use *scriptura continua*. Word-based algorithms to recognize text therefore inevitably fail.

In this paper, I discuss the possibilities of Transkribus (TRANSKRIBUS Team at University of Innsbruck 2019), a program that makes uses of machine learning algorithms and that allows for training one's own models to transcribe texts regardless of language and script style, as long as a sufficient amount of manually transcribed training data are available. The paper is structured as follows: First, I give a brief overview of previous attempts to scan handwritten texts. Afterwards, I briefly present the functionality of Transkribus and the underlying technology. Following that, I discuss the models that I trained using the Transkribus HTR+ engine and

---

<sup>1</sup> I would like to thank Stefanie Anemüller, Andreas Groo, Insa Klemme, Zaidan Lahjouji, Martin Meindl, Lora Taseva and Eckhard Weiher as well as the Transkribus staff, particularly Johanna Walcher and Tobias Hodel for providing valuable help. The usual disclaimers apply.

analyze their real-life performance. In doing so, I touch upon the issue of specificity versus universality with regard to model performance. I conclude with an outlook on the implications of the capabilities of Transkribus for editorial practice in particular and Paleoslavica studies in general.

### **State of the art**

OCR (Optical Character Recognition) for text in printed modern standard language can be considered a mainly solved task with character error rates well below 1%. However, for different stages of historical texts, this does not hold. Early printed books and, even more severely, manuscripts, pose considerable difficulties because of heterogeneity, often mediocre image quality and lack of standardization.

Recently, however, OCR training for early printed books and Church Slavonic editions considerably improved. Projects such as the edition of the Bdinski Sbornik ([bdinski.obdurodon.org](http://bdinski.obdurodon.org)) or the RRuDi corpus (<http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new/>) successfully used OCR technology for digitalization of printed edition. Regardless, there are still a lot of problems when it comes to scanning handwritten texts. Traditional OCR models fail when being used on manuscripts. The results are considerably worse than with printed text and mostly practically unusable.

However, research on recognizing handwritten text in general and Old Cyrillic handwritten text in particular gradually gained traction. As early as 2008, Kornienko et al. proposed to make use of artificial intelligence approaches using neural network technologies for text recognition of Old Slavonic manuscripts and early printed books.<sup>2</sup> Overall, under ideal circumstances, they reached up to 80% correctly recognized glyphs<sup>3</sup> for traditional Slavic manuscripts (Корниенко et al. 2008). While being impressive for the time of conducting the project and a considerable step forward as compared to previous approaches, the results have greater theoretical than practical value, since with an error rate of 20% of all characters at best, more or less every single word would require manual correction. This, in turn, would slow down the correction process to a level that renders the automatic pre-processing step impractical.

However, a free (but in the future freemium) platform called Transkribus has been available for some time. Transkribus is a web- and server-based GUI tool for transcribing manuscripts and handwritten text. Its most interesting feature is the possibility to train models for transcribing

---

<sup>2</sup> There is a large and growing body of literature on artificial neural networks. For a first overview, the reader is referred to Inzaugarat 2018.

<sup>3</sup> According to other information, the real-world performance of their tool was up to 70%.

manuscript text. The text recognition technology implemented in Transkribus is also based on artificial intelligence and neural networks. As opposed to traditional OCR approaches, it uses the so-called HTR approach. According to the Transkribus FAQ page, “Unlike OCR, HTR does not focus on individual letters. Instead, it scans and processes the image of entire lines and tries to decode this data” (Questions and Answers – Transkribus Wiki 2018). Recently, a new algorithm based on advanced neural network technologies called HTR+ was developed and implemented in Transkribus. As compared to the traditional HTR algorithm, it significantly shortens training time<sup>4</sup> while, at the same time, improving the accuracy of text recognition. The experiments reported in this paper are based on the new HTR+ algorithm.

### **Training models with Transkribus**

Training a neural network model is an example of supervised machine learning. This means that, in order to successfully train the model, one needs a specific amount of manually corrected training data. With respect to training handwritten text recognition, one, thus, needs digital images of the manuscript to be transcribed and a diplomatic edition that exactly complies with the manuscript. As a general rule, the more training data are available, the more accurate the model will be. According to the Transkribus FAQs, a minimum of 15,000 transcribed words is recommended in order to produce a usable model. Since neural network models do not rely on linguistic knowledge, the language or alphabet the manuscripts are written in do not play a role. Comparing the characters, words, and lines of the transcription with the respective digital images, the model learns the correct digital transcription of handwritten glyphs. In doing so, numerous epochs of comparing the predicted transcription results with the correct data are needed. The more often the model sees the transcribed data, the better the model adapts to the specific handwriting style of the sources. In Transkribus, the default value for training HTR+ models is 200 epochs.<sup>5</sup> A typical example of the learning curve of a Transkribus model can be seen in Figure 1:

---

<sup>4</sup> The longest training time for a model used for this paper was around 14 hours.

<sup>5</sup> For further information on the specific workflow, the reader is referred to the Transkribus tutorials available online. The terminology related to training neural networks is discussed in Brownlee 2018.

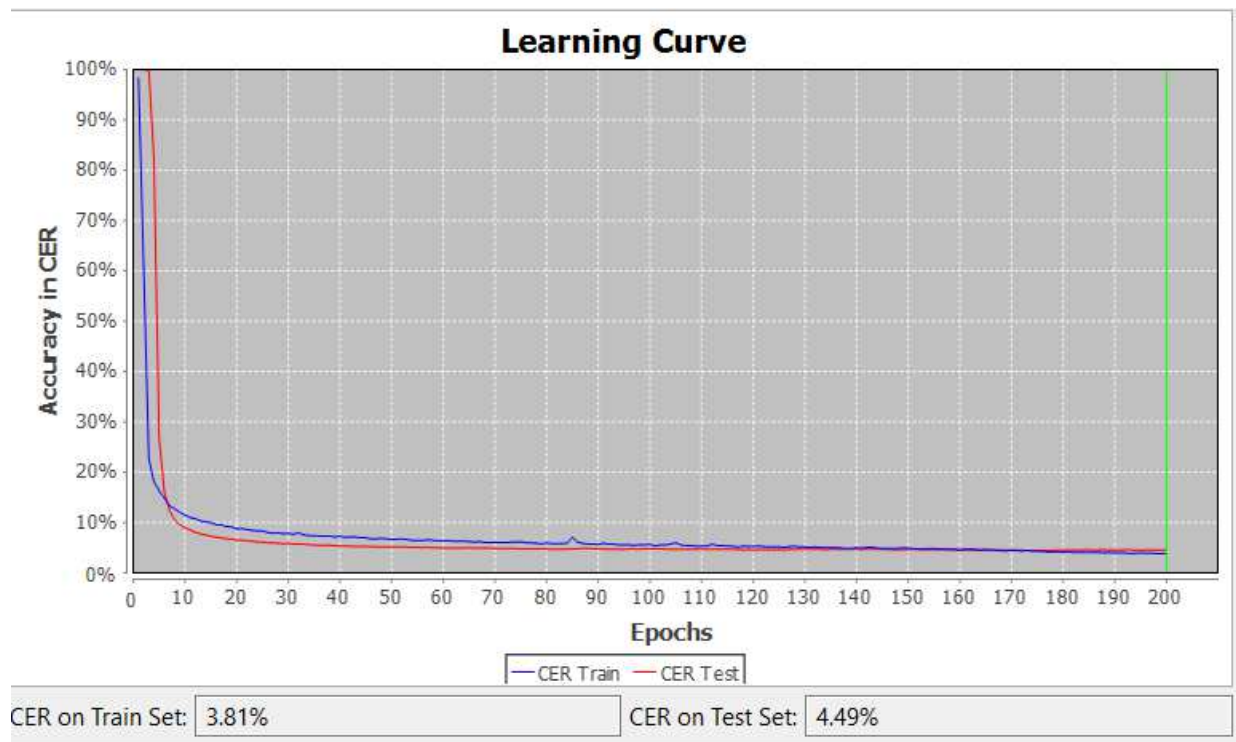


Figure 1: Example of the learning curve of a Transkribus HTR+ model

The computed accuracy is measured in CER (Character Error Rate). The model depicted in Figure 1, trained on a combination of parts of the Codex Suprasliensis and GIM 478 (both being manuscripts from before the 12<sup>th</sup> century written in different types of Cyrillic *ustav*), has a CER of 3.81% on the train set and of 4.49% on the test set after 200 epochs of training. In machine learning, the performance of a model is usually evaluated on a test set, which is part of the overall data, but has not been seen during training. In the specific case of training handwriting recognition models with Transkribus, certain pages of the manuscript(s) used for model training are set aside as a test set. Naturally, the performance using the train set (i.e. data seen during training) is usually better than the performance using the test set (i.e. data not seen during training). As one can see, during the initial 10 or so epochs, CER drops drastically with both training and test data. After that, CER drops less and less significantly with each additional epoch. A typical neural network model training curve has, thus, a hyperbolic shape.

Using many epochs typically leads to the asymptotic approximation of the training data CER curve to zero, which means that the model has more or less completely adapted to the training data. However, while the fact that using a great number of epochs during training takes up many computational resources is getting less and less significant nowadays due to improved hard- and software capabilities, another important unintended effect brought about by using many

training epochs must not be neglected. Since more training epochs adapt the model to the *training* data better and better, it may happen that model performance on *test* data stagnates or even becomes worse. This phenomenon is called *overfitting* and is quite common when training neural networks. Figure 2 shows an example of overfitting using parts of Codex Suprasliensis training and test data.

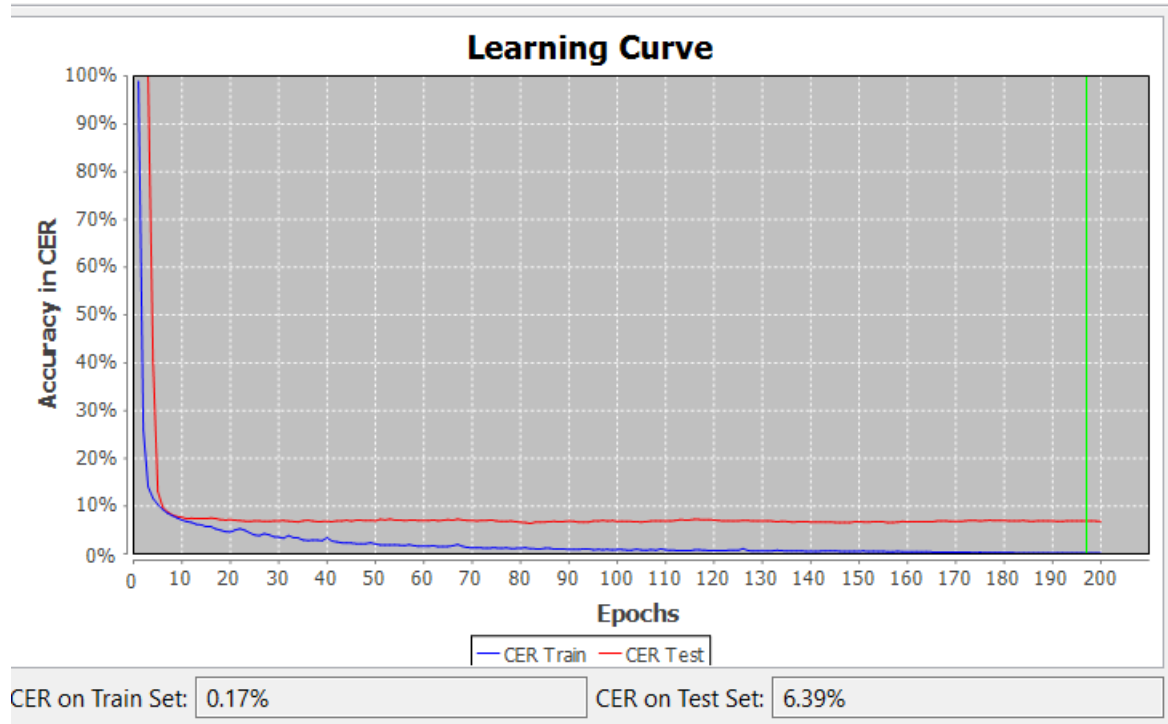


Figure 2: Example of overfitting a HTR+ model

As one can see, starting from around the tenth epoch, CER on the test set does not get significantly better, whereas train CER continuously improves. Since performance on (unseen) test data is the relevant index for real-world transcribing performance with actual new manuscript data, an important goal during model training has to be to avoid extreme cases of overfitting. One method to do so is Early Stopping (Skalski 2018), i.e. one has to find the number of epochs where training and test CER begin to diverge significantly or where test CER increases again, and then stop model training.

In the current implementation of the Transkribus software, there is no viable option to stop model training early after certain conditions – such as the permanent increase of test CER – are met. Instead, one has to resort to ex-post analysis, visually inspect the output graphs and train a new model with just as many epochs as suggested by the test CER curve.

Traditional diplomatic editions of Slavic manuscripts can – and should – be used to train models for transcribing other manuscripts of a similar handwriting. Since many diplomatic

transcriptions of manuscripts are already available in digital form (be it as digital editions on the internet or as word processor files on local disks used for publishing printed books), a plethora of potential training data for different Slavic writing styles are available. For the experiments discussed in this paper, I used parts of the VMČ (Russian Church Slavonic, *poluustav*, 16<sup>th</sup> century), the Codex Suprasliensis (OCS, *ustav*, 11<sup>th</sup> century), and the Catecheses of Cyril of Jerusalem according to a manuscript dating back to the 11<sup>th</sup> century for both training and testing purposes. Other manuscripts used for real-world testing of model performance are mentioned below.

### **Pre-processing and upload**

Some of the texts used for training HTR+ models had to be converted from legacy encoding to Unicode prior to processing in Transkribus. Although Transkribus theoretically works with any encoding scheme, it is clearly advisable to use Unicode data for training purposes. Using legacy non-Unicode data for training would lead to new documents being transcribed using these models in non-Unicode encoding, which isn't recommended for a series of reasons. For the conversion of legacy documents to Unicode I used the conversion tool presented in Skilevic 2013.<sup>6</sup> I did not alter the transcription principles of the individual editions, though. Because of that, some heterogeneity with respect to the rendition of diacritic signs, glyph variants, superscript characters and editorial addenda such as hyphens at line ends is attested. This has to be taken into account when assessing the performance of the combined models.

In order to map the transcribed text to the correct image regions, one needs to identify text regions and add baselines to the images. Transkribus has a function to conduct that task automatically. The exact workflow is described in detail in the Transkribus tutorial. Manual post-processing includes, i.a., removing incorrectly recognized text regions or baselines, e.g., modern additions such as paginations with Arabic figures written with pencil and added by 19<sup>th</sup> century librarians.

---

<sup>6</sup> The first implementation of a new legacy encoding conversion routine takes some time, since the mappings of the respective non-Unicode value to the respective Unicode correspondence have to be entered manually. If one wishes to convert several documents that use the same legacy font/encoding, it will definitely be worth the effort.

## Models of individual manuscripts

### VMČ

The first model I trained with the Transkribus HTR+ engine consists of the Russian Church Slavonic edition of the Apostolos in the version of the *Velikie Minei Čet'i* (Besters-Dilger et al. 2014) and the respective digital images. The features of the model can be found in Table 1:

Name	Words	Lines	Epochs	Train CER	Test CER
VMČ1	173287	38374	200	3.72%	3.82%

Table 1: Features of model VMČ1

As can be seen, the CER is below 4% with both the train and the test set. This is, on paper, an impressive result, according to which, using this model, less than one in 25 letters will be recognized incorrectly. For real-life performance, particularly relevant to Paleoslavists, I provide some qualitative examples<sup>7</sup>. The first example is from f. 999v of the VMČ manuscript used for training the model VMČ1. This folio has not been seen during training.

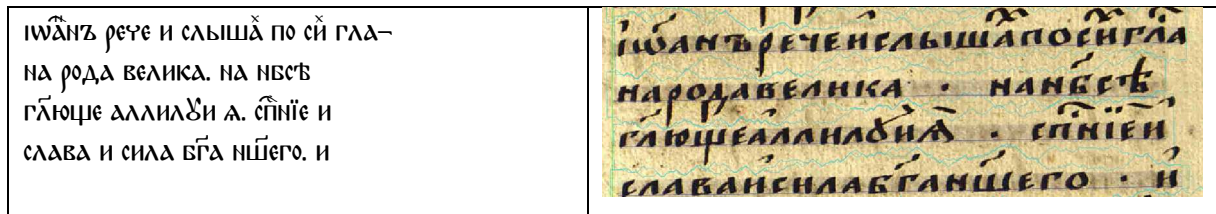


Figure 3: Performance of VMČ1 on SIN 995, f. 999v, lines 3ff.

Here, all errors that occur during automatic transcription are related to diacritic signs, superscript characters and word separation. All simple (i.e. non-superscript or diacritic) characters are recognized correctly, though. The incorrect rendition  $\text{I}\text{W}\text{A}\text{N}\text{Z}$  instead of  $\text{I}\text{W}\text{A}\text{N}\text{Z}$  is interesting, because apparently the model has seen the variant with superscript " and, because of that, tries to reproduce this form despite the fact that there is no sign of a superscript " in the digital image besides the titlo that could be confused with superscript characters. Separating a presumed preposition  $\text{NA}$  in  $\text{NA}$   $\text{PODA}$ , albeit wrong, is not the dumbest error that a machine could commit from a philological point of view.

Transcribing another section of the manuscript (f. 1000r, line 6ff.), the model does not exhibit any issues with respect to superscript characters, titla, and word separation:

<sup>7</sup> The turquoise lines around the letters in the images provided below are produced by Transkribus during the transcription process.



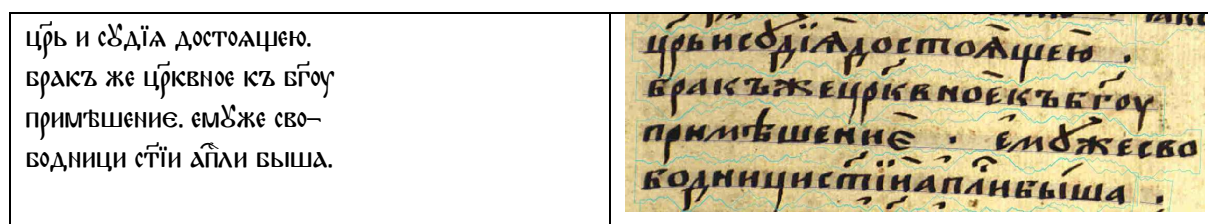


Figure 4: Performance of VMČ1 on SIN 995, f. 1000r

Here, all diacritic marks, all superscript characters, hyphens and blank spaces are set correctly.<sup>8</sup> The real-world performance demonstrated in Figure 3 and Figure 4 is, in my opinion, rather convincing, meaning that the model can be successfully implemented into a transcription workflow. While the cited passages haven't been seen during training, they come from the same manuscript as the training data and, thus, do not allow for generalizations. A different manuscript written in similar script type is the famous Gennadian Bible (Sin. 915, John 3, 13f.), a section of which is used for another real-life evaluation of the model presented in Figure 5:

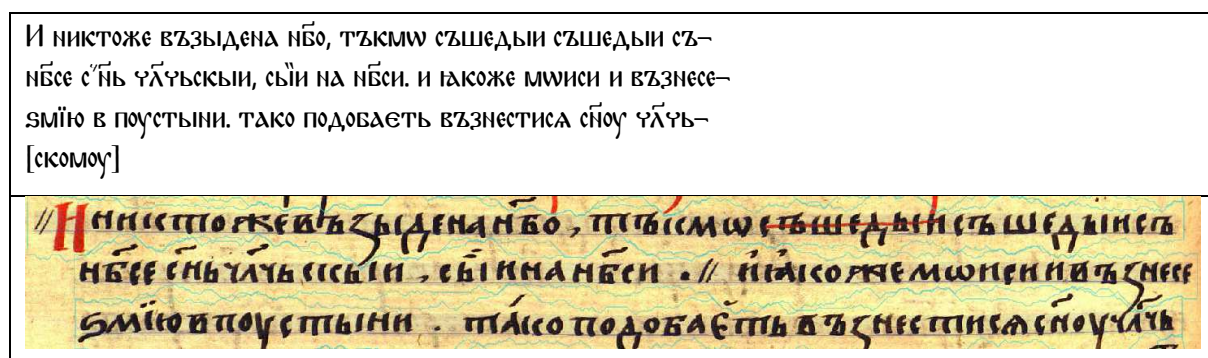


Figure 5: Performance of VMČ1 on SIN 915, John 3, 13f.

The performance is similar to that encountered when using data from the VMČ manuscript as test data: All basic characters are recognized correctly.<sup>9</sup> Problems arise with diacritic signs (сѣнь), superfluous hyphenation<sup>10</sup> and word separation (възыдена, мѣиши и). Still, the results are definitely usable. Manual correction of automatically transcribed data with an error rate similar to the examples provided can be done in a reasonable amount of time and presumably faster than manually transcribing the passage from scratch.

In order to assess the significance of training epochs and training data for model performance, I trained another model with more training data and twice as many epochs:

<sup>8</sup> It has to be noted that the editors of the training data resolved to ignore pneumata, acute, and gravis. The model, thus, could not learn these signs from the given training data and, consequently, ignored them.

<sup>9</sup> In the VMČ training data, 3 and ꙗ have not been distinguished, which is why they cannot be distinguished in HTR using that model.

<sup>10</sup> The reason why the model tends to overproduce hyphens at line endings may be the fact that the VMČ training data are written in two columns. Short lines make line breaks within words more likely than long lines.



Name	Words	Lines	Epochs	Train CER	Test CER
VMČ2	252179	53576	400	3.53%	3.91%

Table 2: Features of model VMČ2

Interestingly, using more data leads to a slightly (albeit statistically not significantly) worse test CER. This may be due to a non-representative selection of test data. We can state that, with homogeneous data of the amount available in our experiments, doubling the training epochs does not lead to a significantly better model.

The real-world results of the model VMČ2 are as follows:

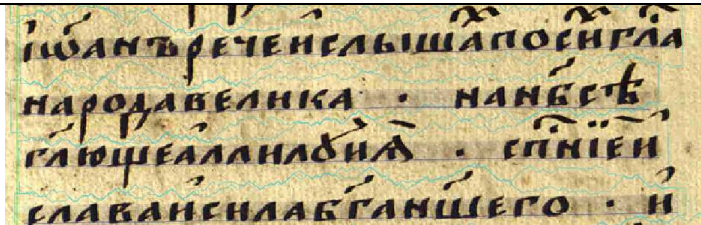
<p>ІВАНЪ РЕЧЕ И СЛЫША ПО СѦ ГЛА  НАРОДА ВЕЛИКА. НА НБСѢ  ГЛЮЩЕ АЛЛАДИИ А. СПНІЕ И  СЛАВА И СИЛА БГА НШЕГО. И</p>	
--	--

Figure 6: Performance of VMČ2 on SIN 995, f. 999v

Interestingly, as opposed to VMČ1, the model VMČ2 renders the phrase гла народа correctly. It also adds the correct titlo to нбсѢ; however, it hypercorrectly adds a superscript <sup>х</sup>, which is quite understandable, since the plural form нбсѢ is a valid Church Slavonic token that has apparently often been seen during training.

The second example from the VMČ manuscript is transcribed using VMČ2 as follows:

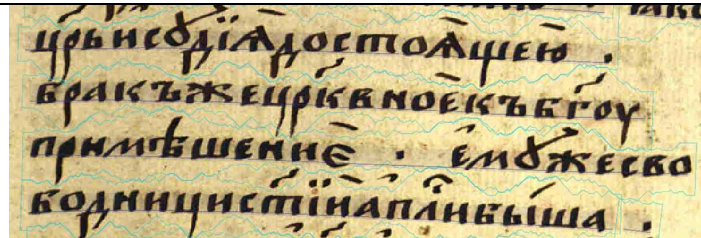
<p>ЦРЬ И СЪДІА ДОСТОАЩЕЮ.  БРАКЪ ЖЕ ЦРКВНОЕКЪ БГОУ  ПРИМѢШЕНИЕ. ЕМОЖЕ СВО-  БОДНИЦИ СТИИ АПЛИ БЫША</p>	
--	--

Figure 7: Performance of VMČ2 on SIN 995, f. 1000r

As opposed to VMČ1 that transcribed this passage without any errors, here, the model fails in correctly rendering both the *titlo* and the blank space in црквноекъ. As one can see from these examples, the real-life performance of models with similar computed CER varies in a way not always predictable. This is confirmed when assessing the performance of VMČ2 on the Gennadian Bible:

<p>И НИКТОЖЕ ВЪЗЫДЕ НА НБО, ТЪКМВЪ СЪШЕДЫИ СЪШЕДЫИ СЪ-  НБСЕ С НЬ УЛЪСКИИ, СЫИ НА НБСИ. И ТАКОЖЕ МВИСИ И ВЪЗНЕСЕ  СМІЮ В ПОУСТЫНИ. ТАКО ПОДОБАЕТЪ ВЪЗНЕСТИСА СНОУ УЛЪ-</p>
--

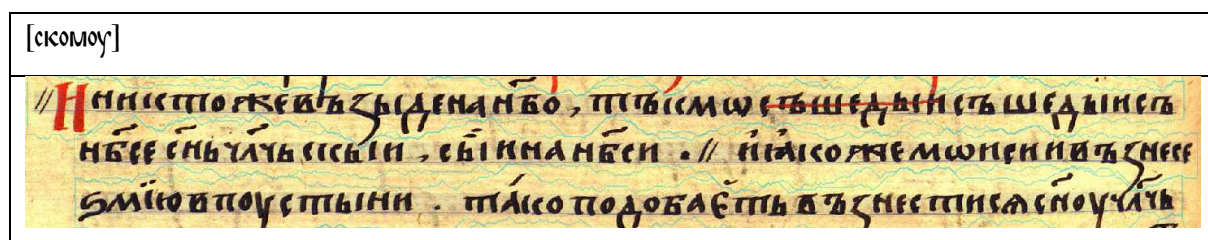


Figure 8: Performance of VMČ2 on SIN 915, John 3, 13f.

с ѣъ is still wrong, but arguably slightly better than сѣъ produced by VMČ1. Furthermore, as opposed to VMČ1, VMČ2 correctly recognizes that възнесе смію are two separate words that do not need to be connected with a hyphen.

Overall, both VMČ1 and VMČ2 perform quite well, while producing errors at slightly different text passages. For optimized results in a transcription workflow, one could think about having both models transcribe the target manuscript independently and visually inspect and manually correct merely those passages where the transcriptions diverge. This approach would be the digital version of the double-key approach (two human transcription teams transcribe one text independently) traditionally often applied in retro-digitalization. It might be worthwhile to further follow this approach in future studies.

In order to explore the minimal amount of training data needed to produce usable results, one model with less than 6000 tokens of training data was trained:

Name	Words	Lines	Epochs	Train CER	Test CER
VMČ_low	5698	1266	200	0.11%	7.28%

Table 3: Features of model VMČ\_low

Judging from the quite significant discrepancy between test and train CER it is obvious that this model is overfitted. Moreover, as compared to the other models, the computed test CER of 7.28% is not very impressive.

The following is an example of the performance of VMČ\_low:

сванъ рече и слыша вси гла- на рода велнка. на нбѣхъ глаголюща иже иже. спитъ и слава и сила бѣа нѣшего. и	сванъ рече и слыша вси гла- на рода велнка. на нбѣхъ глаголюща иже иже. спитъ и слава и сила бѣа нѣшего. и
---	---

Figure 9: Performance of VMČ\_low on SIN 995, f. 999v

While certain remarkably good recognition results can be seen (cf. the last line or the non-spacing of иа), it is clear that the overall performance of this low-resource model is considerably worse than that of the models VMČ1 and VMČ2 trained with lots of data (discrimination of и

and n, superscript letters, confusion of simple letters, etc.). Thus, when faced with the choice of using an already existing model trained with large amounts of data or manually creating a small amount of new training data, one is advised to choose the former, provided the handwriting that is to be transcribed is very similar to the data the respective model was trained on.

## Cyril of Jerusalem

The second manuscript serving as training data is the oldest extant manuscript containing the Slavic translation of the Catecheses of Cyril of Jerusalem, edited by Weiher (2017), Sin. 478, probably dating back to the 11<sup>th</sup> century. This manuscript is written in clean *ustav*. However, the parchment of the manuscript is quite dark, reducing the contrast with the ink. The model trained with the full (minus a couple of test pages) diplomatically transcribed data – having been converted to Unicode prior to being used for training purposes – has the following characteristics:

Name	Words	Lines	Epochs	Train CER	Test CER
Cyrill	66613	9494	200	2.37%	4.83%

Table 4: Features of model Cyrill

Judging from the difference between train and test CER, this model seems to be slightly overfitted. Real-life performance of Cyrill is as follows (fol. 20v, 1 ff.):

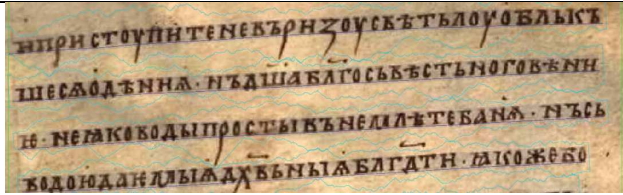
и пристѹпите не въ ризѹ сѣтълоу ѡблыкѹ ше са одѣниѧ. нъ дѣла бѣгосѣвѣстнаго вѣни- е. не ꙗко воды просты въземаѣте бана. нъ съ водою даѣмыа дѣхъныа блгдѣти. ꙗкоже во	
---	--

Figure 10: Performance of Cyrill on Sin. 478, f. 20v

Word separation is the main problem here as well: ѡблыкѹ ше has to be ѡблыкѹ-ше, and бѣгосѣвѣстнаго вѣни-е is to be rendered as бѣгосѣвѣстнѹ говѣни-е. Besides, errors with superscript dots occur, but this might be due to the training data not being entirely accurate with respect to that feature.

The next sample is from folio 228v, 1 ff.

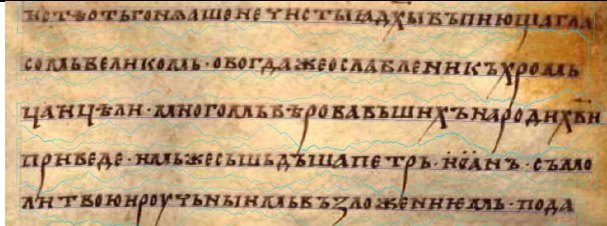
истѣ отыгонаше неустыѧа дѣхъ възпѣющаа глн- сомъ великомъ. овогда же ослабѣниѹ хромъ ца ицѣли. многомъ вѣровавшихъ на роди хѣви приведе. имже съшѣдѣша петръ. и ѡанъ. сѣмо- ли твою и роуѣныимъ възложениемъ. пода- [ста]	
--	--

Figure 11: Performance of Cyrill on Sin. 478, f. 228v

Here we encounter a transcription error committed by the model Cyrill1 with a simple (non-superscript) letter due to low contrast of parchment and ink: *гли-сомъ* should be rendered as *гла-сомъ*. This shows that high contrast is an important prerequisite for successful automatic transcription, an issue that can be addressed in pre-processing of the digital images. Besides, we encounter the already familiar issues with word separation and hyphenation (*хромъ ца* instead of *хромъ-ца*, *на роди* instead of *народи*, *сѣмо-ли твою* instead of *сѣ мо-литвою*). Nevertheless, the model is definitely usable for manuscripts with similar script types.

In the following, the model is tested on other old manuscripts written in *ustav*, namely the Codex Suprasliensis, which has some other paleographic features (f. 9r, 1ff.):

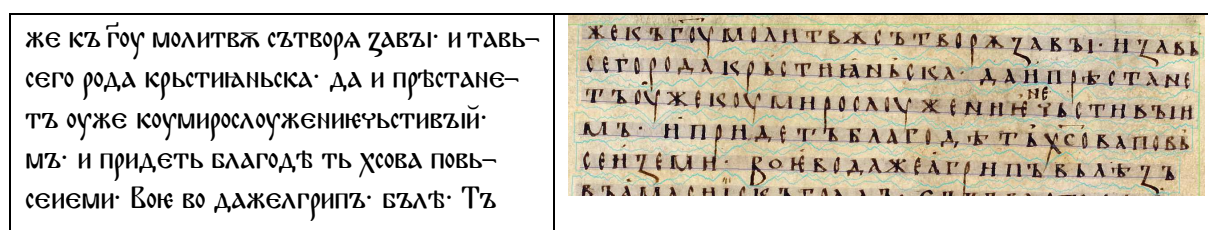


Figure 12: Performance of Cyrill1 on Codex Suprasliensis, f. 9r

The different style of *z* in Suprasliensis as compared with the Cyrill1 training data is clearly noticeable: The model correctly recognizes *z* merely one out of four times. In the other cases, it confuses this letter with small or capital *т* or omits it altogether. Other errors – besides word separation errors – include the confusion of *ж* and *а* as well as *а* and *л*, which is quite understandable from a paleographic point of view.

It has become clear that the performance of a model on manuscripts with different paleographic features, while somewhat usable, is considerably worse than with a manuscript closely resembling the training data manuscript. This becomes even more obvious when using the model VMČ2 trained on *poluustav* script for recognizing the *ustav* script represented in Codex Suprasliensis:

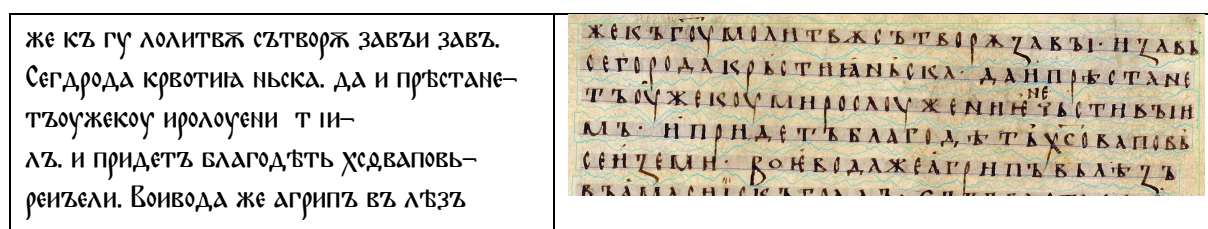


Figure 13: Performance of VMČ2 on Codex Suprasliensis, f. 9r

The broad *ustav* *м* seems to pose specific problems to the model. Besides, some characters such as *ѣ* have not been seen in the 16<sup>th</sup> century training data of VMČ2, which is why they couldn't



be recognized correctly. The combination of several unknown glyph forms and superscript characters led to extremely poor performance (т ии– instead of ѱѣтивзи–). Thus, it becomes clear that we are in need of a model that is more robust towards individual variation, which is why I experiment with a combination of training data from different manuscripts below.

The Ostromir Gospel (Luke, 25, 17ff.) is recognized using the model Cyril1 as follows:

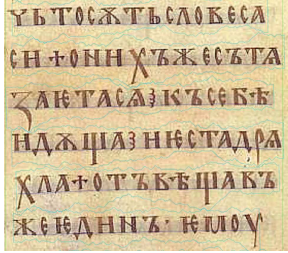
ѱѣто сѣтъ словеса– сито ниѣже сѣта– заѣта са і кѣ себѣ Идѣцаи кѣта дра– хлаотѣ бѣцавѣ же єдинѣ ємоу–	
---	--

Figure 14: Performance of Cyril1 on Ostromir Gospel, Luke, 25, 17ff.

The main issues with the very clear *ustav* of Ostromir Gospel are signs that are no characters such as the cross † that obviously confuse the model and trigger wrong word separations. Another issue is that the model sometimes interprets the letters as minuscule and sometimes as capital letters. This issue is even more severe when using the VMČ1 model on the Ostromir Gospel:

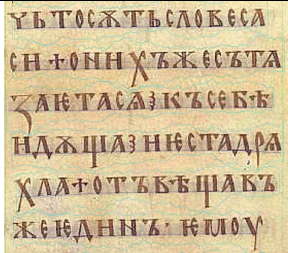
Бѣ то бѣтъсповеса сижо ниѣже сѣта злѣтлоа кѣ себѣ– Идѣца, и истадра– хлакогѣбѣцавѣ же юдинѣ. и тоу–	
--	--

Figure 15: Performance of VMČ1 on Ostromir Gospel, Luke, 25, 17ff.

It is obvious that this result is unusable and far worse than Cyril1, mostly due to the letters such as є unseen during training of VMČ1.

## Codex Suprasliensis

The Old Church Slavonic Codex Suprasliensis is available online in a remarkable digital edition ([suprasliensis.obdurodon.org](http://suprasliensis.obdurodon.org)). I used this edition for training another model based on Old Church Slavonic *ustav* data.

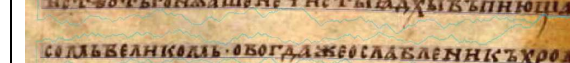
Name	Words	Lines	Epochs	Train CER	Test CER
Suprasl1	71519	11764	200	1.56%	4.29%

Table 5: Features of model Suprasl1


же кѣзъ гоу молитвѣ сътворѣ завѣи ѿ за вѣ-  
сего рода крѣстианѣска. да ѿ прѣста не  
тѣ ѡу же коу мироу служениѣ чѣстивѣи  
мѣ. ѿ придетъ благодѣтъ хѣова по вѣ-  
сен земли. Воіевода же агригъ вѣлѣтъ

Not surprisingly, Supras11 is best adapted to transcribe data from Codex Suprasliensis. Besides the superscript  $\text{ne}$ , in this example there are merely errors with respect to word separation. Using Supras11 for transcribing the other *ustav* manuscript in our sample, Cyril, does not make much sense, as the results are rather poor.

и пристоупите не въризоу свѣтъпоубыль-  
шеслохъ. нъ аша багосвѣтного вѣни-  
е. не іако воль простъвъ немиѣте банѣ. нъ съ-  
водою даіемиѣ вънѣ багати. іакоже бо

<p>ИСТЪ ОТЬГОНАШОЕ НЕЧИСТЪІАХЪІ ВЪ ПИИША СОМЪ ВЕЛИКОМЪ. ОВОГДА ЖЕ ОБЛАВЕНИКЪ АРОА ЖДА ИЦЪІАИ. МНОГОМЪ ВЪТРОВАВЪШІХЪ НАРОДИВИ ПРИВЕДЕ. ИМЪЖЕ СЫШДЪША ПЕТРЬ. И ОАНЪ. СЪМО- ЛИ ТВОЮ ПРІОУЧЪНЫМЪ ВЪЗЛОЖЕНИЕМЪ. ПОРА- [СТА]</p>	 <p>ИСТЪ ОТЬГОНАШОЕ НЕЧИСТЪІАХЪІ ВЪ ПИИША СОМЪ ВЕЛИКОМЪ. ОВОГДА ЖЕ ОБЛАВЕНИКЪ АРОА ЖДА ИЦЪІАИ. МНОГОМЪ ВЪТРОВАВЪШІХЪ НАРОДИВИ ПРИВЕДЕ. ИМЪЖЕ СЫШДЪША ПЕТРЬ. И ОАНЪ. СЪМО- ЛИ ТВОЮ ПРІОУЧЪНЫМЪ ВЪЗЛОЖЕНИЕМЪ. ПОРА-</p>
--	---

Many incorrectly recognized characters are attested here that render Suprasll unusable for transcribing data similar to the handwriting attested in Sin. 478. Besides, while the performance of the model Suprasll with the Ostromir Gospel manuscript isn't spectacular and exhibits numerous errors, an interesting feature occurs in the transcription:

<p>             ѸТОСЖТЬ СЛОВЕСА              СИ ТО НИХЪ ЖЕ СЪТА-              ЗАЕТА САЗЪ СЕБѢ              ИДЖ ШАРИНЕСТА ДРЖ-              ХИЛЪ ОТЪВѢШТАВЪ              ЖЕ ЕДИНЪ. ЕМОУ           </p>	 <p>             ѸТОСЖТЬ СЛОВЕСА              СИ ТО НИХЪ ЖЕ СЪТА              ЗАЕТА САЗЪ СЕБѢ              ИДЖ ШАЗ НИСТАДРА              ХЛАТОТЪВѢШАКЪ              ЖЕ ЕДИНЪ. ЕМОУ           </p>
---	---

14

Instead of  $\delta\tau\zeta\epsilon\upsilon\mu\alpha\epsilon\zeta$ , the model uses  $\delta\tau\zeta\epsilon\upsilon\mu\tau\alpha\epsilon\zeta$ , thus transcribing the ligature  $\mu$  with its individual components  $\mu$  and  $\tau$ . While there is no visual cue for the model triggering the rendition of  $\mu$  as  $\mu\tau$ , the form  $\delta\tau\zeta\epsilon\upsilon\mu\tau\alpha\epsilon\zeta$  has been seen in Suprasl training data several times. Apparently, taking into account what it has learned during training, the model resolved that, judging from the surrounding letters, the most probable characters in the middle of the word in question are  $\mu\tau$ , thus exhibiting some kind of linguistic intelligence.

### Combined models

In order to test whether it is possible to produce generic models that are not specifically adapted towards a certain type of handwriting, I trained a model more or less containing all data from all three sources, i.e. essentially a combination of VMČ, Cyril, and Suprasl.

Name	Words	Lines	Epochs	Train CER	Test CER
Comb1	393079	75422	400	4.42%	3.92%

Table 6: Features of model Comb1

Despite having trained the model for 400 epochs, train CER is still higher than test CER. This suggests that, for fitting models with such a high amount of words taken into account, one could use even more epochs without overfitting the model.

I start with the same qualitative examples as above for evaluating the combined model:

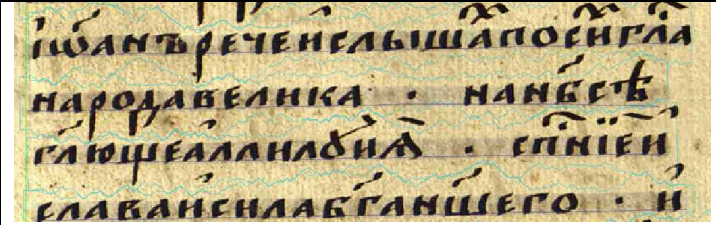
<p>и҃ѡа́нъ рече и слышѧ по сѣ г҃ла— народа велика. на нѣсѣ г҃люще а҃ллилѣи а. с҃пїе и слава и сила б҃га нѣшего. и</p>	
---	--

Figure 20: Performance of Comb1 on Sin. 995, f. 999v

Here, the model performs slightly better than VMČ1 (it correctly recognizes the titlo in нѣсѣ), but somewhat worse than VMČ2 (e.g., it does not correctly recognize г҃ла).

The second example from the VMČ is rendered the following way:

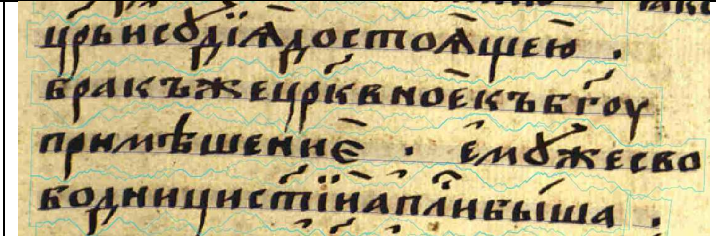
<p>црѣ и сѣдѧ достоѣно. бракѣ же црквиное кѣ б҃гоу примѣшение. емѣже сво— бодници с҃тїи а҃пѣи быша</p>	
--	--

Figure 21: Performance of Comb1 on Sin. 995, f. 1000r



Here, the model's performance is slightly better than VMČ2, but (due to the missing titlo in црквиное) slightly worse than VMČ1.

With respect to the Gennadian Bible, the model renders the text as follows:

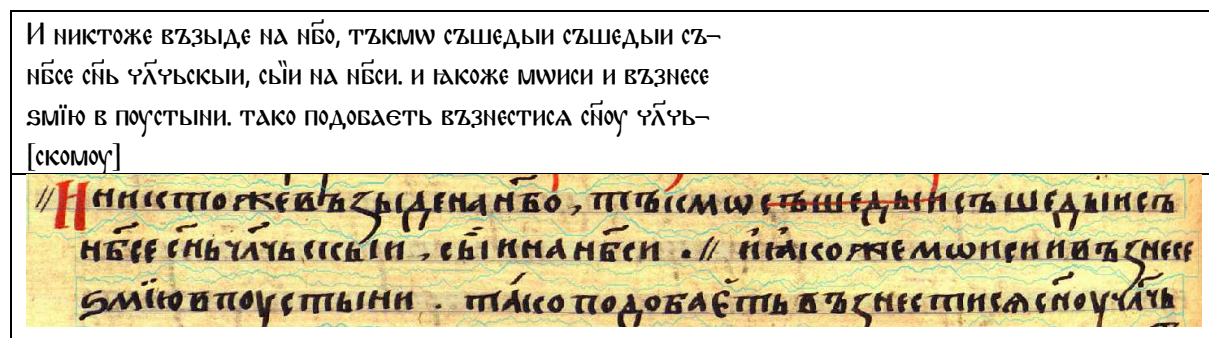


Figure 22: Performance of Comb1 on Sin. 915

Interestingly, as opposed to the specialized models VMČ1 and VMČ2, this model correctly recognizes сѣнь, while still struggling with hyphenation and word separation (съ-нбѢ and мѦИСИ и).

As has been shown, the performance of the combined model Comb1 with *poluustav* is comparable to the performance of specialized models. With respect to *ustav* manuscripts, the situation is as follows:

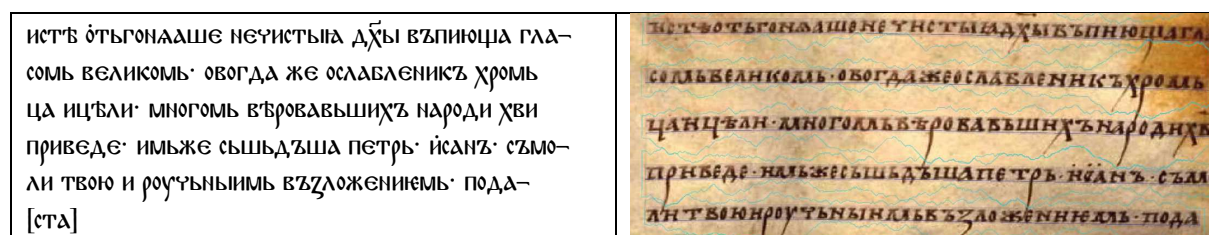


Figure 23: Performance of Comb1 on Sin. 478

Despite the low contrast, гла-сѦ is recognized correctly. Maybe, due to the high amount of training data from different sources, the model has learned that гла-сѦ is a valid Slavic word as opposed to гла-сѦ, the transcription produced by the specialized model Cyril1. Other errors in hyphenation and word separation are similar to the ones produced by the specialized model. With respect to исанъ instead of иѡанъ, the combined model performs worse than the specialized one.

The transcription of the section from Codex Suprasliensis exhibits certain issues, the most significant being the fact that it could not cope with the superscript Ѣ in the third line, which, apparently, confused the model quite significantly:

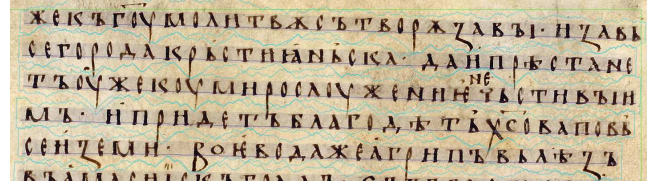
<p>же кѣ гѡу молитвѣ сътворѣ за вѣ· и за вѣ- сего рода крѣстианьска· да и прѣстане- тъ оу же коумиролужениѣ вѣтивъи мъ· и придетъ благодѣтъ хѡва по вѣ- сен земли· Воіевода же агрипъ вѣлѣтъ</p>	
--	--

Figure 24: Performance of Comb1 on Codex Suprasliensis

When using the combined model with completely different scribal styles (manuscript A. Mazurin, f. 36v, thanks are due to A. Miltenova for providing me with parts of the manuscript), the results deteriorate as compared to scribal styles seen during training, but not to the extent that renders them completely unusable.

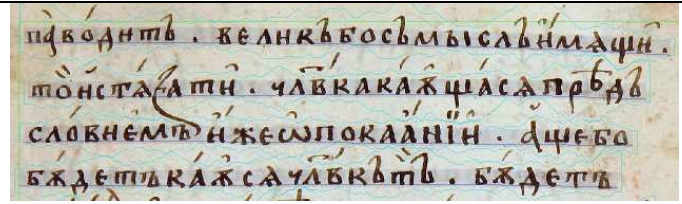
<p>паводитъ. великъ бо съмыслъ имаши. то истазати. члвѣка каѣща прѣдъ- словнемъ иже ѡ покаяніи. аще бо бѣдетъ важа члвкътъ· бѣдетъ</p>	
--	--

Figure 25: Performance of Comb1 on Mazurin

The specific forms of а and к are sometimes recognized correctly, sometimes they are confused with л and в, respectively. The long з is never confused with ѣ, though, and the idiosyncratic з in истазати is recognized more or less correctly. Interestingly, the model chose з and not ꙗ to render that glyph, despite the fact that ꙗ was seen during training numerous times. This may be due to a quantitative bias towards the VMČ training data. As mentioned above, as opposed to the older *ustav* training data, the VMČ training data does not contain ꙗ.

Compared to the combined model Comb1, VMČ1 fails with glyphs not (or very seldom) seen during training such as з and ꙗ when transcribing this manuscript.

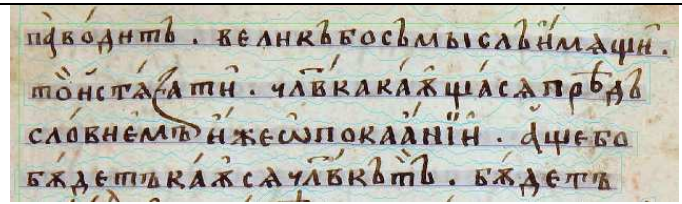
<p>паводитъ. великъ бо съмыслъ имаши. то истазати. члвѣка каѣща прѣдъ- словнемъ иже ѡ покаяніи. аще бо бѣдетъ важа члвкътъ· бѣдетъ</p>	
--	--

Figure 26: Performance of VMČ1 on Mazurin

The model Cyril1 performs as follows with the unseen data:

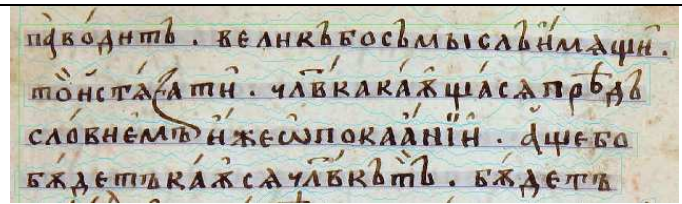
<p>паводипо· великъ бо съмыслъ имаши· по йстаапи· члѣвакаѣща са прѣдо- слови емви же ѡповаани· аще бо бѣдепѣкаж са члѣкъ пѣ· бѣденъ</p>	
---	--

Figure 27: Performance of Cyril1 on Mazurin

It is obvious that the model Cyril1 is next to unusable on this manuscript, due to errors in almost every token. The combined model shows, thus, more flexibility with respect to new scribes and new writing styles.

This becomes also apparent when analyzing the performance on another previously completely unseen manuscript, the Bdinski Sbornik (f. 40r., taken from [bdinski.obdurodon.org](http://bdinski.obdurodon.org), cf. (Birnbaum et al. 2014))

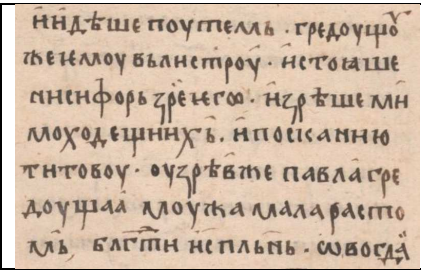
<p>нидѣше поуѣмь. Гредоуѣи же ѣмоу въ листроу. и стоѣше ниифоръ зрѣ его. изрѣшеми мохоещнихъ. и посканню титовоу. оузрѣже павла гре- доуѣа моужа мала расто- мь блѣти испльнь. Ѡвогда</p>	
---	--

Figure 28: Performance of Comb1 on Bdinski

The overall transcription quality is rather good, but apparently, the model Comb1 struggles with discriminating и and н.

VMČ2 performs as follows:

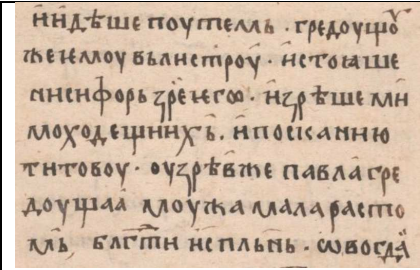
<p>нидѣше поуѣмь. Гредоуѣи же ёмоу въ ли строу. истоѣше ниифоръ зрѣ его. изрѣшеми плохоещнихъ. и посканню хитовоу. оузрѣже павла гре- соуѣаа лоужаллааа расто- гль. блѣти испльнь. Ѡвогда</p>	
---	--

Figure 29: Performance of VMČ1 on Bdinski

It has become clear that the specialized model VMČ performs considerably worse than the combined model Comb1, mainly due to the incorrectly recognized broad м that often triggers further errors.

The case with model Cyril1 is somewhat less obvious. Cyril1 copes remarkably well with the task of discriminating н and и. In this respect, it performs significantly better than Comb1. However, it fails with other discrimination tasks, above all it is unable to recognize the specific form of т, consistently confusing it with н.

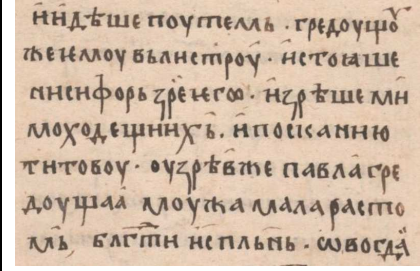
и ѿдѣше поуѣмь · грѣдоуцѣ же ѿмоу вѣлипроу · и стоѣше ниси фортъ зрѣ юго · изрѣшеми мохощицѣхъ · и посканию ти товоу · оузрѣже павла грѣ- доущаамоу · жа мала распо- мь · блги испльнь · ѿвогда-	
--	--

Figure 30: Performance of Cyrilil on Bdinski

Interestingly, the combined model Comb1 fails with the large, clear *ustav* found in Ostromir Gospel:

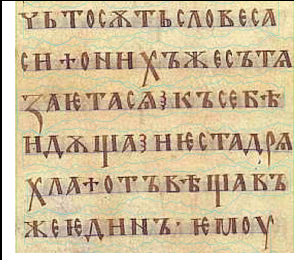
бѣто сѣтъ словеса- сѣто нѣхъже сѣта- заѣта сѣ кѣ сѣбѣ и дѣна и сѣта дѣра- лѣтѣ оуѣтъцавѣ же ѿдѣнъ · ѿмоу	
---	--

Figure 31: Performance of Comb1 on Ostromir Gospel

The mediocre result is mainly – but not exclusively – due to the model unsuccessfully trying to establish a system of using uppercase and lowercase letters in the manuscript. Converting all uppercase letters in lowercase letters during pre-processing of the training data could potentially alleviate this issue.

With the exception of the somewhat unexpectedly poor performance transcribing the Ostromir Gospel, our results convincingly show that combined models such as Comb1 generally cope better with completely new and unseen handwriting styles that do not closely resemble handwriting styles present in the training data. Models such as Comb1 can, thus, efficiently be used for the first draft transcription of manuscripts where there is no training data available at all.

## Conclusion and outlook

Our analysis has shown that handwriting recognition models trained with the HTR+ engine provided by Transkribus produce usable results when applied to Old Slavic manuscripts. Thanks to machine learning, it is now possible for the first time to have the computer produce a first draft of a transcription that, while definitely in need for competent manual correction, can be corrected using less time and money than producing the draft manually. Together with the fuzzy search and keyword spotting capabilities of Transkribus (that should be explored in greater detail in further studies) making it unnecessary to have 100% error-free text



transcriptions in order to successfully find keywords, this allows for the mass digitalization of previously unpublished manuscripts and the use of these texts for philological and, above all, corpus linguistic purposes in the future. HTR is, thus, an important cornerstone in the process of digitalizing our field in a speed previously unknown.

While models specifically trained on one particular handwriting usually perform best, first attempts to create generic models yielded decent results. These generic models, brought about by combining the training data of several specialized models and training the combined data for numerous epochs, have the distinct advantage that they can be successfully applied to manuscripts completely unseen during model training. Paleoslavists interested in experimenting with other manuscripts and the models presented in the current paper, are encouraged to do so. Upon request, I can make these models available to a wider audience via the Transkribus platform.

With respect to creating universally usable generic models, the results can potentially be further improved by adding more, and more diverse, training data from different sources. In order to successfully accomplish this goal, I appeal to all colleagues to share digital images and diplomatic transcriptions that can – and will solely – be used for improving the transcription quality of the models. Joining forces in this respect may eventually lead to a situation where the majority of available Slavic manuscripts can be digitized by mouse click in a short amount of time and with a low error rate. Cooperation in training and using HTR+ models is key, and it will be beneficial for everyone interested in Paleoslavistics.

## **Publication bibliography**

Besters-Dilger, Juliane; Halapats, Viktoria; Kindermann, Natascha; Maier, Elina; Rabus, Achim (Eds.) (2014): *Die großen Lesemenäen des Metropoliten Makaij. Uspenskij Spisok. Kommentierter Apostolos*. München: Kubon & Sagner (Studies on language and culture in Central and Eastern Europe, 22).

Birnbaum, David J.; Dobbeleer, Michel de; Popowycz, Alexandre; Sels, Lara (2014): *Bdinski sbornik*. Available online at <http://bdinski.obdurodon.org/>, checked on 4/15/2019.

Brownlee, Jason (2018): *What is the Difference Between a Batch and an Epoch in a Neural Network?* Available online at <https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/>, checked on 4/17/2019.

Inzaugarat, Eugenia (2018): Understanding Neural Networks: What, How and Why? – Towards Data Science. Available online at <https://towardsdatascience.com/understanding-neural-networks-what-how-and-why-18ec703ebd31>, checked on 4/15/2019.

Questions and Answers – Transkribus Wiki (2018). Available online at [https://transkribus.eu/wiki/index.php/Questions\\_and\\_Answers](https://transkribus.eu/wiki/index.php/Questions_and_Answers), updated on 12/19/2018, checked on 4/16/2019.

Skalski, Piotr (2018): Preventing Deep Neural Network from Overfitting – Towards Data Science. Available online at <https://towardsdatascience.com/preventing-deep-neural-network-from-overfitting-953458db800a>, checked on 4/11/2019.

Skilevic, Simon (2013): SlaVaComp: Konvertierungstool. In *Slověne* 2 (2), pp. 172–183, checked on 4/11/2019.

TRANSKRIBUS Team at University of Innsbruck (2019): Transkribus. Available online at <https://transkribus.eu/Transkribus/>, checked on 4/15/2019.

Weier, Eckhard (Ed.) (2017): Die altbulgarische Übersetzung der Katechesen Kyrills von Jerusalem. Freiburg i. Br.: Weier (Monumenta linguae slavicae dialecti veteris, tom. 64).

Корниенко, С. И.; Черепанов, Ф. М.; Ясницкий, Л. Н. (2008): Распознавание текстов рукописных и старопечатных книг на основе нейросетевых технологий. Available online at <https://textualheritage.org/ru/el-manuscript-08-/52.html>, checked on 4/15/2019.

**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 2018 the President of the Commission. 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.