# An OCR system for the Unified Northern Alphabet

Niko Partanen
Institute for the Languages of Finland
niko.partanen@kotus.fi

Michael Rießler
Bielefeld University
michael.riessler@uni-bielefeld.de

**Abstract**

This paper presents experiments done in order to build a functional OCR model for the Unified Northern Alphabet. This writing system was used between 1931 and 1937 for 16 (Uralic and non-Uralic) minority languages spoken in the Soviet Union. The character accuracy of the developed model reaches more than 98% and clearly shows cross-linguistic applicability. The tests described here therefore also include general guidelines for the amount of training data needed to bootstrap an OCR system under similar conditions.

**Tiivistelmä**

Tutkimus esittelee Yhteiselle pohjoiselle aakkostolle kehitettävään tekstintunnistusmalliin tähtääviä kokeita. Kyseistä aakkostoa käytettiin 16 Neuvostoliiton pohjoiselle kielelle noin vuosina 1931–1937. Kehitetty malli saavuttaa merkkitasolla parhaimmillaan yli 98% tunnistustarkkuuden, ja se kykenee tunnistamaan samalla kirjoitusjärjestelmällä kirjoitettuja eri kieliä. Tehtyjen kokeiden perusteella tehdään arvioita siitä, kuinka suuria aineistomääriä tarvitaan uuden tekstintunnistusjärjestelmän toteuttamiseen.

## 1 Introduction

This article describes the tests conducted recently as part of the Kone Foundation-funded IKDP-2 project on developing an OCR system for the Unified Northern Alphabet, a writing system used during a period of time for several languages spoken in Northern areas of the Soviet Union. Part of the work has been conducted in the Institute for the Languages of Finland in relation to the OCR and HTR experiments recently carried out at the institute. The study uses openly available materials so that the resources created and evaluated here can be used further in downstream NLP tasks. The trained models and the scripts used to create them, alongside the evaluation scripts, are all published alongside the paper as an independent data package





*Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages*, pages 77–89,
Tartu, Estonia, January 7 - January 8, 2019.

(Partanen and Rießler, 2018a)[1], which allows them to be fine tuned and used directly for text written in this writing system.

OCR systems are known to be well suited for specific writing systems and fonts rather than specific languages. Adding more sophisticated language models to known OCR systems has proven challenging, especially for agglutinative languages. For instance, trying to integrate a morphological analyser into the Tesseract system produced much worse results than using simple wordlists (Silfverberg and Rueter, 2014). The extent to which OCR systems require linguistic information is unclear, which raises question of how language-independent they even are. The research question in our study emerges from this issue, and we evaluate whether an OCR system developed using multiple languages performs better or worse than a system that has seen data from only one specific language. The remaining errors will be analyzed separately in order to shed more light on the current bottlenecks in the model.

Another important question is to what extent it is possible to bootstrap an OCR system for new, rare or idiosyncratic writing systems. In the course of their histories, Uralic languages have seen a very large variety of such writing systems, and many writing conventions are used only in a small number of publications. It is therefore important to have a general understanding of the amount of resources needed to reach results that are comparable to the current state of the art in OCR applications. Only with this information is it possible to decide on which tasks further work should focus on. The goal of this kind of work is not necessarily to develop an OCR system that works for one language in a wider set of situations, but simply to extract the texts of individual publications so that they can be used for linguistic research purposes, in addition to other applications.

Reul et al. (2018) describe how the various OCR systems have shifted to use recurrent neural networks, which results in typical character error rates (CER) below 1% for books published using modern typographic conventions. Early printed books show more variation and may often require book-specific training to reach CERs below 1–2%. Since the Soviet publications from the 1930s presumably qualify as non-modern prints, these figures also provide a baseline for our study.

## 2 History of the Unified Northern Alphabet

The Unified Northern Alphabet (UNA) was developed for 16 minority languages of Northern Russia in the late 1920s and taken into use in 1930. It is connected to the Latinization process in the Soviet Union, which started during the early 1920s and was first introduced to Islamic populations that had previously used the Arabic script (Grenoble, 2003, 49). In the 1930s, the alphabet was extended to cover more languages, including several very small languages for which UNA became the first common writing standard in 1932. In principle, UNA is similar to other Latin alphabets created during the same period. For the smaller northern languages, UNA represented the first effort to create an alphabet, whereas for other languages the Latin scripts replaced the systems that had previously been in use.

UNA works on the principle that all languages use the same base forms of characters, which are modified with diacritics depending from the phonological requirements of individual languages. The system seems to have been used in a phonologically consistent manner, so that the characters chosen for each language represent the phonetic realization of phoneme in the given language.

[1] https://github.com/langdoc/iwclul2019

The languages for which UNA was used are listed below (cf. (Siegl and Rießler, 2015, 203)), with ISO 639-3 codes in parentheses:

- Aleut (ale)
- Central Siberian Yupik (ess)
- Chukchi (ckt)
- Even (eve)
- Evenki (evn)
- Itelmen (itl)
- Ket (ket)
- Kildin Saami (sjd)
- Koryak (kpy)
- Nanai (gld)
- Nivkh (niv)
- Northern Khanty (kca)
- Northern Mansi (mns)
- Northern Selkup (sel)
- Tundra Nenets (yrk)
- Udege (ude)

In connection with this process, a large number of textbooks and dictionaries were published (Grenoble, 2003, 164). Since these books were printed in St. Petersburg and clearly designed using common materials, they are very close to one another in their content and style. The fact that these materials were intended to be used in creating literacy among these peoples explains why there are no translations of the same books in larger languages of the Soviet Union, which also had their own widely translated titles.

UNA was abandoned in 1937 in favour of individual Cyrillic writing systems. In practice, this change halted the written use of these languages for decades to come, and the next written standards did not arise until the 1950s, or much later in the case of certain languages (Siegl and Rießler, 2015, 204–205).

It is unknown to us how many books were ever published in UNA, but based on searches in various library catalogues, the number is probably some dozens per language. This is not an enormously large corpus, but it is still enough that for languages that have extremely narrow resources at the moment, the digital accessibility of these resources can be of utmost importance. The fact that these books are starting to be old enough to be released as Public Domain even further increases their value. Already the fact that these books can be used for any purposes without licensing issues should speak on behalf of their wider inclusion in different corpora.

Texts published in UNA are also very important for the current documentation efforts, since they represent the language as it was used almost a century ago. It is clear these texts have their drawbacks and represent only a limited range of genres, but still they certainly complement the other types of resources very well and are worth further research.

## 3 Materials used

A large number of books written in UNA is available in the Public Domain as part of the Fenno-Ugrica collection (The National Library of Finland, 2018).[2] In addition to this, individual texts can be found in the collections of other libraries.

P. N. Žulev's primer was translated to several languages using UNA, and the Kildin Saami (Žuļov, 1934), Northern Mansi (Žuļov, 1933a), Northern Selkup (Žuļov, 1934) and Tundra Nenets (Žuļov, 1933b) versions are available in Fenno-Ugrica. In addition

[2] `https://fennougrica.kansalliskirjasto.fi/`


(a) Kildin Saami

(b) Northern Mansi

(c) Northern Selkup

(d) Tundra Nenets

Figure 1: Examples from P. N. Žulev's primer in various languages

to this, the E-resource repository of the University of Latvia offers an Evenki version of the primer (Zulew, 1933).

The first Ground Truth package for UNA was recently published (Partanen and Rießler, 2018b) by the authors of this paper. This is essentially a collection of manually corrected lines in the different languages. Our study uses a sample from the version 1.1 of the package, which is available in GitHub.[3]

Figure 1 illustrates the way the alphabet was used, showing matching excerpts from P. N. Žulev's primer. The texts are not completely identical translations in each language. The content differs to some degree, for instance for various culture-specific backgrounds. The translations have also been published separately in Russian, which indicates that the differences may be more significant. For example, there is a Russian translation of the Mansi primer[4], and similar Russian editions exist for other languages too. This is a clear sign that they are not only translations from one source. To our knowledge, no analysis of these differences has been conducted.

The Figure 2 displays the cover image and page 5 of the Evenki version of the book. The font is the same as before, but obviously the language is different, with some

[3] https://github.com/langdoc/unified-northern-alphabet-ocr
[4] http://urn.fi/URN:NBN:fi-fe2014060526307


(a) Cover

**Ɜəlgərə, əkəl damgatira.**

Ɜəlgərə əcə əwiɜərə. Ɜəlgərə onŋandu damgatiɜaran.

Ɜalunca gunən:

— Əkəl damgatira. Ənunkə oɜaŋas. Oktaʙrcikar, tatigawɜəril kuŋakar əwkil damgatira.

Kuŋakar gunə:

— Ɜəlgərə, nodakəl mogdiwi. Əməkəl, əwilgət.

**Bərkəwul tawicimni.**

Ɜalunca Bərkəwultiki gunən:

— Oldoksowo təsikəl. Bagdamawa dukuŋkiwa nəkəl. Taduk kuŋakardu dukuwurwatin, dukuŋkilwatin ʙukəl. Əsi-tirga si tawicimni ʙisinni. Mit upkat amarulta tawitcarap.

**Ajicimni.**

Ajicimni tatkittula əmərən. Nuŋan kuŋakar ŋaləlwətin minərən. Lurɜawul ŋələllən. Tuksamurən. Ajicimni gunən:



5

(b) Page 5, cf. figure 1

Figure 2: Cover and page 5 of P. N. Žulev's Evenki primer


Figure 3: Kildin Saami title in capital letters

new characters. The usual font in the books is shown in figures 1 and 2. The books also contain some headlines that are in very unusual all-caps typeface that is only sporadically covered in the train and test data. Figure 3 illustrates this. Since these lines are very rare, they are not used in our experiments at all. The Ground Truth package metadata is used to distinguish these lines. Since they occur only in very specific portion of the book, the Ground Truth package does not contain examples of this in all four languages.

However, the texts still exhibit some variation. For example, some elements are in bold font, and these are kept as they are. They are not separately tagged in the Ground Truth data either, although one could suggest this as an improvement so the effect of the presence of different font types in the training and testing sets as well as the accuracy rate for these font types could be better evaluated.

To contextualize further what kind of data this is, these books contains on average 100 words per page, the number of characters being on average 600–700 per page. The lines, of which there are usually 20 per page, have around 30 characters on average. One page contains approximately 100 words, with great variation depending on image locations and spacing around titles. These numbers are not exact since they represent only the Ground Truth data, which does not contain the whole content of the books. Still, the figures are similar across the translations and can be seen as highly representative.

## 4 Experiment design

The model training is done with Ocropy (Breuel, 2008)[5], as it offers a very convenient set of tools for various OCR-related tasks. Other options would have been Tesseract and Transkribus, and repeating the tests with various systems should be carried out in further research.

Ocropy, as with other modern OCR systems, is given training data as pairs of line images and corresponding text. The text recognition is distinct from Layout Analysis, which refers to element detection and line segmentation, with the goal of finding the lines in their correct order. It is important to note that when we speak of OCR accuracy we mean the accuracy for already correctly segmented lines. The model is given line-based material, which Ocropy keeps learning iteratively, saving the model at regular intervals. The number of iterations controls the time the model is given to train. The model learns the correspondence of line images and texts, and it does not need any specific font or character style information. If a character does not exist in Unicode, as is the case with several letters used in UNA, a mapping has been done in Ground Truth to visually similar but factually incorrect letters. This is done simply to aid visual inspection of the results, as mapping could have been done for any unique characters.

The primary languages involved in the study are Kildin Saami, Northern Selkup, Tundra Nenets and Northern Mansi. These were used in the Ground Truth package, and the large amount of Kildin Saami material made it possible to design our study so

[5] https://github.com/tmbdev/ocropy

that Kildin Saami could be compared to a setting in which all four languages are mixed together in the same OCR system. The third Evenki experiment is also explained below.

The Ground Truth package was sampled and processed for our experiments with a script that prepares the working environment for the experiments. It is provided with other documentation in an additional data packagePartanen and Rießler (2018a) stored in a GitHub repository associated with the paper. The repository also contains detailed examples of how to reproduce all plots and figures presented in this study.

In the first experiment, the idea was to test the amount resources needed to bootstrap an OCR system in this kind of situation. We tested the training of a model on different amounts of lines, divided equally into subsets that are equivalent to pages (an addition of 20 lines counted as an increase of one page). Twenty experiments were carried out, for an incrementally growing amount of training material. The Ocropy system was trained for 10,000 iterations per model.

In the next experiment, two different OCR models were trained using a larger, apparently sufficiently sized, body of training material. One model was trained on all four languages in equal proportions, and the other with only data from Kildin Saami. In this experiment, the model was trained for 50,000 iterations and the number of training lines was also larger, 200 lines per language, for a total of 800 lines. Similarly, the Kildin Saami monolingual model was trained for an equal number of iterations and with 800 lines.

The test sets common for both experiments contained 100 lines per language, or altogether 400 lines. A test set that is half the size of the training set may seem too large, but this seemed reasonable since otherwise the number of lines in individual languages would have been so small that it would have been uncertain whether the different characters were at all equally present. Similarly, one of our primary topics of investigation was whether a practical OCR system could be built with these resources and training scenarios, which makes extensive testing reasonable.

Since we aim to provide an OCR system for the Unified Northern Alphabet, it would be important to test the system on a language that is not at all included in the current models. This would truly reveal whether the OCR system actually generalizes toward the whole writing system. With this in mind, the Evenki dataset described in section 3 was used as an additional test experiment. The scores on the Evenki dataset were reported and analysed in context, but this data was not used in training in any of the models.

Section 6 contains an error analysis. In this section, the error output of Ocropy is evaluated in order to identify the language-specific bottlenecks that keep the error rate high in some test scenarios.

# 5 Results

## 5.1 Gradual page increase test

Figure 4 shows the gradual improvement in the accuracy of the Kildin Saami model as the number of training pages is increased. The figure shows that the model improves very quickly when more pages are added for training. With 8 pages, the model reaches an error rate approaching 2%, and falls below that if the number of pages is increased to 11. The remaining mistakes are analysed further in section 6. By increasing the training time per model and adjusting other parameters, this accuracy could

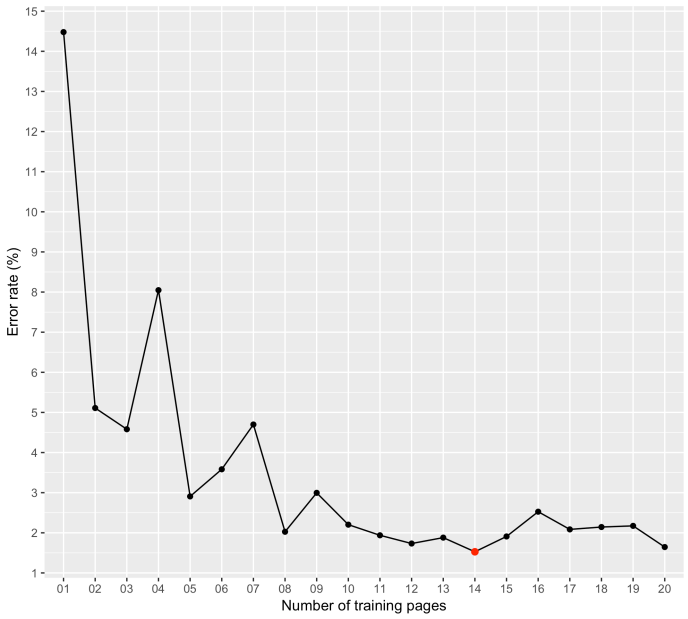


Figure 4: Test scores for Kildin Saami OCR model. Best score with error rate of 1.527% with 14 pages marked with red

maybe have been reached even earlier, but the increase in the amount of training data clearly brings continuous improvements in accuracy. In itself this is not surprising, and nothing else could have been expected from this experiment.

However, the test does offer some very valuable insight. After 5 pages, the error rate had already fallen into 2.91%. This is perhaps not yet a state-of-the-art level, but a character accuracy of 97% is already rather effortless and quick to proofread. Individual percentages can be squeezed out by increasing the number of pages, but in order to OCR an entirely new book, five pages, or approximately 100 lines, seems to be enough to bootstrap a useful OCR system that, although not necessarily ready for production, can at least be used to produce the needed increase in the number of pages more quickly and easily.

## 5.2 Comparable monolingual–multilingual test

This test aims to compare the performance of OCR models trained using monolingual and multilingual materials on different language specific-test sets. Figure 5 follows the pattern observed in the earlier test, as the Kildin Saami reached the same accuracy below 2% that it had also exhibited before. For the sake of clarity, the character sizes and accuracy of the test sets are presented in detail in table 5 and visualized in figure 5.

The Kildin Saami model does not perform equally well on other language tests, which makes sense, since the Kildin Saami model alone has never seen some of the special characters used in these languages. The result with Northern Mansi is the closest, and indeed the difference between the Northern Mansi character set and that of Kildin Saami is also the smallest. The errors are more thoroughly discussed in section 6.

The mixed test does not outperform the monolingual Kildin Saami model, which,

| model | test | errors | characters | error percent |
|---|---|---|---|---|
| mixed | mns | 45 | 3428 | 1.313 |
| mixed | sel | 31 | 3772 | 0.822 |
| mixed | sjd | 61 | 3405 | 1.791 |
| mixed | yrk | 60 | 3564 | 1.684 |
| sjd | mns | 110 | 3428 | 3.209 |
| sjd | sel | 421 | 3772 | 11.161 |
| sjd | sjd | 54 | 3405 | 1.586 |
| sjd | yrk | 442 | 3564 | 12.402 |

Table 1: Mixed and monolingual OCR models compared

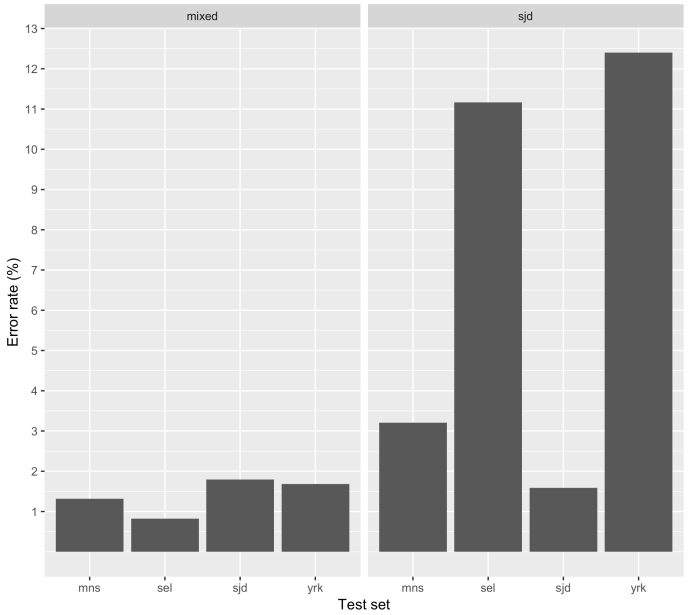


Figure 5: Mixed and monolingual OCR models compared

indeed, has had only one fourth of the exposure to the Kildin Saami special characters that the monolingual Kildin Saami model received. Nevertheless, the results are very close. Even more importantly, the mixed model achieves above 98% accuracy for all of the four languages, and above 99% accuracy for Northern Selkup. The experiment demonstrates that from the point of view of one language, it does not make a very big difference whether the 800 lines used in training are from the same language or from four different languages, as long as the character set is shared.

### 5.3 Additional Evenki test

The Evenki test was conducted using the same model as in the previous test presented in section 5.2. The error rate was 5.073 % using the mixed model and 12.832 % using the Kildin Saami model. This falls well below the accuracy of the previous tests but is in line with the early phases of the gradual page increase test. Important conclusions can also be drawn from the fact that the Kildin Saami result is close to the Kildin Saami results on Selkup and Tundra Nenets – Evenki is equally foreign to the Kildin model as these languages are, as would be expected.

## 6 Analysis and error evaluation

Some of the characters recognized poorly belong to a group of characters that generally resemble one another quite a lot; especially pairs such as I : l, e : є, s : ѕ, z : ƶ are confused occasionally even with the best-performing models. A more common type of remaining mistake comes from uppercase letters. However, since the training has been done in a low-resource scenario with a smaller amount of training data than would be common, the prevalence of capital letters in the errors seems easily explainable. Uppercase letters are used rarely in most of the texts, making up only slightly more than 5% of all letter characters in the training data. From this point of view, it seems obvious that the accuracy of uppercase letters will trail behind the rest until the entire training set has reached a relatively large size.

The Kildin Saami model performed relatively poorly on Selkup and Tundra Nenets. The previous error analysis in this section showed that this was related to the lack of recognition accuracy in those letters that are present in those languages but not in Kildin Saami. The fact that the Kildin Saami model performed rather well on Northern Mansi must be related to the somewhat small character inventory used in Mansi and to the fact that it largely overlaps with the inventory of Kildin Saami. The only character present in Northern Mansi but missing in Kildin Saami, at least in this dataset, is ӈ.

The additional Evenki test further and more profoundly illustrates the problems seen in section 5.2 when testing different language combinations. For example, the Evenki letters that were not recognized by the mixed model were ʒ and ə̄, both of which are rare or non-existent in the current training data. Kildin Saami has four instances of ʒ in entire Ground Truth package in word internal positions, whereas in Evenki this is a highly common character. The Kildin model has a more narrow character set in use than the mixed model, which is illustrated by the very common error that occurs when using Kildin Saami model for Evenki: w : vv. Kildin Saami does not use w in UNA, whereas Evenki does not use v. These differences, when added up, provide a good ex-

planation for the accuracy rates seen in the experiment. They also illustrate how a cross-linguistic writing system such as UNA benefits specifically from mixed language training, as the model has the opportunity to see characters across the languages.

A further type of error comes from numerals, which are very rare in the Ground Truth package. They occur a few times in running text, but at the moment the models simply do not recognize them at all.

## 7 Conclusions

The error rates using mixed model for all languages were below 2%, for Northern Selkup even below 1%. In section 3, we mentioned that one page had on average 600-700 characters. These error rates would translate into 6–12 errors per page on average. The error analysis in section 6 demonstrated that the errors are rather concentrated to specific character pairs.

One observation that arises from our work is that training an OCR model for a new writing system, even with incomplete Unicode support, can be done very easily with the current technology. Arbitrary mapping of line texts and images is, as explained in section 4, in principle independent from whether the characters recognized actually correspond to those that are printed. A fast iterative process where the first model is trained using a very small dataset, which is then used to create a somewhat larger dataset with which the same procedure is repeated, appears to be a very effective and effortless method. Based on our incremental page test, five pages (100 lines) was enough to bring the accuracy up to more than 97% percent, suggesting that the initial model should already be trained with a very small amount of training data, if the situation is indeed such that the training has to be started from scratch. This rate of accuracy results in around 20 corrections per page, which is arguably a bearable task. Our study also indicates that in a situation where there is training material available for some languages, we can use that to train an OCR system that also works sufficiently well on other unseen languages, at least if the entire character set of the target language is covered in the training materials.

The accuracy problems were clearly connected to characters missing from the training data but encountered in test languages, and this is an area where cross-lingual OCR will inevitably experience problems. Uppercase characters were also recognized at a poorer rate than others throughout the tests, and this is obviously connected to their sparsity in the training data. It is difficult to imagine a way around this problem, which is a major bottleneck in low-resource scenarios. One suggestion could be to make sure that even the rarer letters are at least sporadically present in the training data, perhaps by picking out lines in the available materials that contain these characters in initial position. The initial character issue brings even more problems in multilingual scenarios, as there are many language-specific phonotactic limitations on which characters can occur in initial position and will thus be present in uppercase form. Naturally they can still occur occasionally in lines that are entirely capitalized, an instance of which was presented in figure 3. Further research should evaluate whether lines in all capitals improve the accuracy of word-initial capital letters as well.

The use of various languages to train one OCR system provides a potential answer to the question on the degree to which OCR models are language specific and how much they actually generalize across languages. We do not claim that our experiments would have yet shed much light on this question, but further experiments with Unified Northern Alphabet are a good avenue for studying this topic further. For sake of comparison, some scenarios that are similar to OCR recognition of UNA include recognizing texts written in UPA, IPA or Americanist Phonetic Notation. In all these cases, a writing system that is in principle uniform is used across different languages.

Moving forward, full parallel texts should be extracted from these books using the OCR models provided. This data should also be converted into the contemporary orthographies, after which it could be used for a variety of purposes. For example, creating new treebanks within the Universal Dependencies project could be a very interesting way to improve the digital infrastructure of these languages rather visibly. Similarly, language documentation projects working with the endangered northern Eurasian languages should certainly be interested in resources such as texts written in UNA. Since these materials are largely in the Public Domain, there are exceptionally few limitations to what could be done.

## Acknowledgments

The authors collaborate within the project “Language Documentation meets Language Technology: the Next Step in the Description of Komi” funded by Kone Foundation. Thanks to Alexandra Kellner for proofreading this article.

## References

Thomas M Breuel. 2008. The OCRopus open source OCR system. In *Document Recognition and Retrieval XV*. International Society for Optics and Photonics, volume 6815, page 68150F.

Lenore A. Grenoble. 2003. *Language policy in the Soviet Union*. Kluwer Academic Publishers.

Niko Partanen and Michael Rießler. 2018a. An OCR system for the Unified Northern Alphabet – data package https://doi.org/10.5281/zenodo.2506880.

Niko Partanen and Michael Rießler. 2018b. Unified Northern Alphabet OCR Ground Truth v1.1 https://doi.org/10.5281/zenodo.2443922.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. State of the art optical character recognition of 19th century fraktur scripts using open source engines. *arXiv preprint arXiv:1810.03436* .

Florian Siegl and Michael Rießler. 2015. Uneven steps to literacy. In Heiko F. Marten, Michael Rießler, Janne Saarikivi, and Reetta Toivanen, editors, *Cultural and linguistic minorities in the Russian Federation and the European*

*Union*, Springer, number 13 in Multilingual Education, pages 189–229. https://doi.org/10.1007/978-3-319-10455-3$_8$.

Miikka Silfverberg and Jack Rueter. 2014. Can morphological analyzers improve the quality of optical character recognition? In *Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages*. http://dx.doi.org/10.7557/5.3467.

The National Library of Finland. 2018. Fenno-ugrica. `https://fennougrica.kansalliskirjasto.fi/`. Accessed: 2018-12-21.

P. N. Zulew. 1933. Taŋin ʒarin dukuwun: nonopti hanin, nonopti anŋani alagun ʒarin. *E-resource repository of the University of Latvia. Prof. Pētera Šmita kolekcija* https://dspace.lu.lv/dspace/handle/7/28252.

P. N. Ƶuļov. 1933a. Lovintan maʃьs lovintanut: oul lomt :oul haniꙉ tan tal maʃьs. *Fenno-Ugrica collection. The National Library of Finland* http://urn.fi/URN:NBN:fi-fe2014060426213.

P. N. Ƶuļov. 1933b. Tolaŋgowa jeʃjemŋa tolaŋgobçʃj: ŋurtej peļa: ŋurtej toholambawa po jeʃjemŋa padawь. *Fenno-Ugrica collection. The National Library of Finland* http://urn.fi/URN:NBN:fi-fe2014061629286.

P. N. Ƶuļov. 1934. Kniga logkəm guejka: vəsmus pieļ vəsmus egest opnuvmus. *Fenno-Ugrica collection. The National Library of Finland* http://urn.fi/URN:NBN:fi-fe2016051212324.