

## Komi Latin Letters, Degrees of UNICODE Facilitation

Jack Rueter, Larisa Ponomareva

University of Helsinki, Digital Humanities  
jack.rueter@helsinki.fi, dojegpl@gmail.com

### Abstract

The Komi-Permyak and Komi-Zyrian language forms are two writing traditions of the pluricentric Komi language, whose first debut in as a literary medium stems from the 14th Century. The alphabets and writing systems have changed numerous times over the centuries, and most published media in the two languages are facilitated by UNICODE. There is, however, one specific time period, where the digital accessibility of plentiful publications is hampered by missing representation in UNICODE, and that is the time span 1932–1938; the Komi languages were printed in a Latin script with numerous and special Latin letters. Due to the poor quality of typography in this era, the characters have been determined transitional, and therefore it has been suggested that characters existing in UNICODE, regardless of range, might be used to alleviate the issue of missing representation. Bearing this in mind, the number of required letters dropped from an original estimation of 18 to 8 or perhaps simply combining descenders. "UNICODE Script Ad Hoc" has provided helpful suggestions for providing digital means for the Komi languages, and with UNICODE in place, the next hurdle will be dealing with font issues. **Keywords:** Komi Latin alphabet, UNICODE, proposal of new characters, digital accessibility, minority languages

### Дженьта висьталом

Перем коми кыв да зырянскӧй коми кыв – уна быдкодь местаэзын олӧсь коми кывлӧн гижан кыввез. Коми кылын медодззайсь литературнӧй артыс аркмис XIV-ӧт векӧ. Кад съборна коми анбуррез да гижан арттэз унаис вежсьывлӧс. Унажык пассэз, кӧдна пантасьлӧны кык коми кыв вылын лэдзӧм литератураын, эмӧсь ЮНИКОД-ын. Дзир ӧтик кадӧ, 1932-ӧт восянь 1938-ӧт воӧдз, коми кыввез вылын лэдзӧм литература съӧкыта шедӧ тӧдмавны компьютерӧн, сӧдз кыз сӧя кадся комиӧн гижӧмись не быд пас эм ЮНИКОД-ын. Эта кадӧ коми кыввез вылын литература вӧли лэдзӧм латинскӧй шыпассэзӧн, кӧдна коласын вӧли уна быдкодь ассяма латинскӧй шыпас. Сы кадся типографика эз вӧв бур. Эта увья эна пассэз вылӧ пондӧс видзӧтны, кыз вуджан кадся пассэз вылӧ. Этасянь вӧли шум, что пассэз, кӧдна пантасьлӧны ЮНИКОД-ын быдкодь анбуррезын, вермасӧ босьтсьыны тырмытӧм пас мыччалӧм понда. Эта увья колана шыпассэзлӧн лыдыс чинис медодззайсь висьталом 18-сянь 10-ӧдз. Эта вермис лоны сӧдзжӧ пассэзлӧн ӧтлаӧтӧмсянь. "UNICODE Script Ad Hoc" («Юникод Пассэзын торья пантасьӧммез») чукӧр висьталӧс коланаторрез коми кыввез компьютерӧн тӧдмалӧм понда. ЮНИКОД-ӧ колана пассэз пыртӧм бӧрсянь пыкӧт лоас шрифттэзын, шыпассэзлӧн неӧтнӧжа гижӧмын.

## 1. Background

In 2011, a "Language Programme" was drafted at the Kone Foundation in Helsinki, Finland with the objective of promoting research work on endangered languages through funding of individual research projects and the preservation of irreplaceable data sets i.a. In 2012–2013, a Digitization Pilot Project of Kindred Languages<sup>1</sup> was coordinated by Jussi-Pekka Hakkarainen in the auspices of the National Library of Finland to save 1920 and 1930 publications from possible water damage. In the pilot, the National Library of Finland in collaboration with the National Library of Russian in St. Petersburg and with funding from the Kone Foundation "Language Programme" digitized newspapers and school books representative of Balto-Finnic, Mordvin and Mari minority languages. The materials were made available in the Fenno-Ugrica collection at the National Library of Finland<sup>2</sup>

In 2014–2016, the Digitization Project of Kindred

Languages<sup>3</sup> was extended to address publications especially representative of the time span 1932–1937 for the Permic, Ob-Ugric and Samoyedic languages. Due to the express time span 1932–1937, it soon became apparent that little of the Komi publications could, in fact, be totally digitized for UNICODE-based access. Some of the letters were entirely missing. In fact, Komi was not the only one lacking a complete alphabet, but this was also the situation for some of the other minorities of the north, who had a special Latin alphabet developed for them called the Unified Northern Alphabet (Siegel and Rießler, 2015). Naturally, encoding is not the only matter to be dealt with when digitizing text materials. The characters encoded in UNICODE may also require special glyphs for a given language, which makes it possible to capture a typographically uniform set of data. And it might be argued that language specific word lists and morphology could have an effect on recognition (Silfverberg and Rueter, 2014; Partanen and Rießler, 2019).

<sup>1</sup>[https://www.doria.fi/bitstream/handle/10024/94581/Sukukielten%20digitointiprojekti\\_loppuraportti.pdf?sequence=2&isAllowed=y](https://www.doria.fi/bitstream/handle/10024/94581/Sukukielten%20digitointiprojekti_loppuraportti.pdf?sequence=2&isAllowed=y)

<sup>2</sup><https://fennougrica.kansalliskirjasto.fi>

<sup>3</sup>[https://www.doria.fi/bitstream/handle/10024/130799/Sukukieltendigitointiprojekti\\_kansalliskirjasto\\_Hakkarainen\\_FINAL.pdf?sequence=2&isAllowed=y](https://www.doria.fi/bitstream/handle/10024/130799/Sukukieltendigitointiprojekti_kansalliskirjasto_Hakkarainen_FINAL.pdf?sequence=2&isAllowed=y)

## 2. Komi and UNICODE

Komi is a member of the Permic branch of the Uralic language family. It is spoken in the Komi Republic, the Perm Krai as well as parts of western Siberia and the Kola Peninsula. Komi consists of a continuum of dialects represented by two modern literary language traditions: Komi-Permyak and Komi-Zyrian. Although most published materials written in the various Komi alphabets are digitally available through UNICODE, there is a time span (1932–1937), when Latinization co-occurred with prolific publication activities, and access to these texts is hampered in digital spheres by the absence of necessary character encoding.

Since an earlier proposal made for an entire extension block to encode Latin letters used in the Former Soviet Union<sup>4</sup> had not been accepted, it was decided that a new language-specific proposal be made from an entirely descriptive perspective. It was important that the description of the missing letters be concise, so as not to be met with an under-informed evaluation of look-alike characters already present in the Latin range of UNICODE. On such look-alike letter can be in the Cyrillic soft sign <ь>, the Latin tone six <̂>, and the Latin letter <b>, presented in Figure 1.

	U+044A	U+044C	U+0185	U+0062	U+042A	U+042C	U+0184	U+0042
TIMES NEW ROMAN	Ь	ь	б	б	Ь	ь	Ь	B
ARIAL UNICODE	Ь	ь	б	б	Ь	ь	Ь	B
GENTIUM	Ь	ь	б	б	Ь	ь	Ь	B
CALBRI	Ь	ь	б	б	Ь	ь	Ь	B
COURIER	Ь	ь	б	б	Ь	ь	Ь	B
LUCIDA GRANDE	Ь	ь	б	б	Ь	ь	Ь	B
	hard sign	soft sign	tone 6	b	hard sign	soft sign	tone 6	B

Figure 1: Comparing Cyrillic letters soft sign with Cyrillic letters hard sign and Latin look-alikes

### 2.1. Komi Alphabets with UNICODE Support

The Komi literary traditions are first attested in the form of Old Permic script in 1372. Old Permic scripts are attributed to Stefan Khrap (Saint Stephen of Perm) (1340-1396). Attestation of their use date into the 17th century.



Figure 2: Old Permic rendered in UNICODE 10350–1037F.

From the time of the original Old Permic scripts in 1372 to the present, The Komi language forms have been written in three different alphabet ranges:

1372–1600s in Old Permic (Figure 2), 1600s–Present Russian-related Cyrillic, 1918–1938 Molodtsov Cyrillic (Figure 3), interim 1932–1937 transitional Latin (Figure 4).

While use of the Molodtsov Alphabet is shown with a maximal span of twenty years, it was not used in all publications for the whole time period. In fact, publications made in the Permski Krai did not start using the Molodtsov Alphabet until 1921, and then they abandoned it almost entirely in 1932 for a Komi Latin alphabet.



Figure 3: Cyrillic supplement for Molodtsov in UNICODE 0500–050F.

### 2.2. Komi Alphabets without UNICODE Support

The Komi Latin alphabet (Figure 4) (1932–1937) consisted of 36<sup>5</sup> letters. Just like the Molodtsov Alphabet, it was a phonemic system, where each character represented an individual phoneme of the Komi languages.

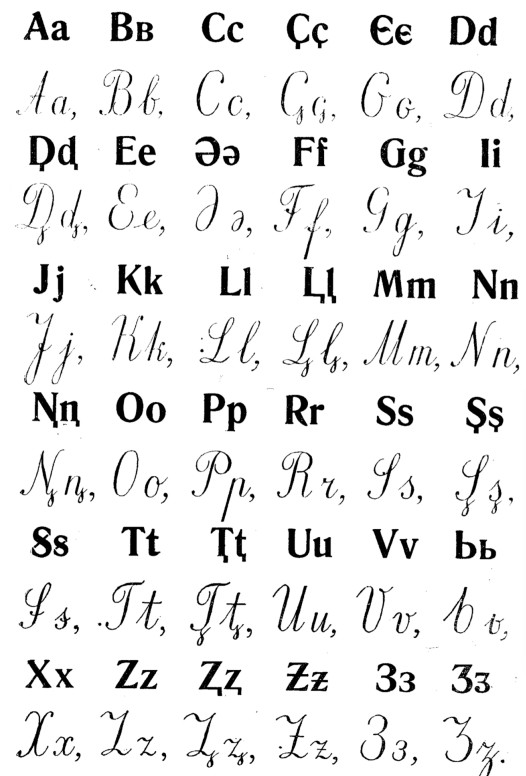


Figure 4: Komi Latin letters A–Z.

The Latin alphabet was used in the Permski Krai from 1932–1937, where it was nearly exclusively used in

<sup>4</sup><http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4162.pdf>

<sup>5</sup>The letters 3з are also rendered as 3з, which increases the number of letters 38.

the publication of both books and newspapers. In the Komi Republic, however, the conversion to the Latin alphabet was never complete. Although it was used systematically in the publication of school books in the Komi Republic (1932–1936), the use of Komi Latin in newspapers varied from mere newspaper names and bolded titles to actual texts. As such falling back on the Molodtsov Alphabet in the Komi Republic (1936–1938) introduced no radical change for many of the readers.

The Komi Latin Alphabet shares many features of the Unified Northern Alphabet (1931–1937) (Partanen and Rießler, 2019) – and both of these character sets were targeted in a previous proposal to the UNICODE Consortium, mentioned above. The Unified Northern Alphabet, in turn, was an attempt to address phonemic features of under-studied languages of the north, and whose Cyrillicization also presented difficulties (Grenoble, 2003).

### 3. Proposal of New Characters

Originally, it was noted that 18 characters were missing from the Latin Range of UNICODE, and therefore a proposal was made<sup>6</sup> in early June, 2019. No immediate acceptance was expected, but relatively prompt response did open new points of departure. There was something we had overlooked in understanding UNICODE principles.

#### 3.1. Thumb Sketch of UNICODE Principles

A proposal for new characters should address a concrete issue. A given alphabet should derive its letters from a single range. And no precomposed letters should be sought when they can be created utilizing combining diacritics from the (+U0300) section of UNICODE or supplements thereto.

#### 3.2. How We Proceeded

We delimited our proposal to the polycentric Komi language and the digital inaccessibility of over half a decade of literary texts. Here we received support from both the National Library of Finland, Finnish Localization (Kotoistus), and FU-Lab in Syktyvkar, Komi Republic, Russian Federation<sup>7</sup>.

We were able to find 27 x 2 letters from within the Latin range, as shown in Table 1.

Aa	Cc	Dd	Ee	Əə	Ff	Gg	Ii	Jj
Kk	Ll	Mm	Nn	Ŋŋ	Oo	Pp	Rr	Ss
Ss	Tt	Uu	Vv	Xx	Zz	Zz	Zz	Зз

Table 1: Letters available in Latin Range

<sup>6</sup><https://www.unicode.org/L2/L2019/19224-n5101-komi-latin.pdf>

<sup>7</sup>The Finno-Ugric Laboratory for Support of the Electronic Representation of Regional Languages <https://fu-lab.ru/laboratoriya>

These 27 letters, upper and lower case, are readily supported in the Latin range of UNICODE. But this means an addition 10 x 2 letters are required, hence the 18 character proposal.

#### 3.3. After Feedback

Within two months of submission, we received feedback on our proposal. It was noted that the typographical quality of books and newspapers printed in the Komi Latin alphabet were not of high quality – many of the composed letters show symptoms of vacillation between descender glyphs. It was determined that Komi Latin was, in fact, a transitional alphabet, and therefore it should not be delimited to a single range.

In removing the range requirement, we were immediately given access to 6 x 2 additional letters, see Table 2. It should be noted, however, that these are only typographical solutions, i.e. the Cyrillic Ukrainian IE (+U0404), (+U0454) <Єє> is used to represent a non-palatalized voiceless coronal affricate, on the one hand, and Cyrillic VE (+U0412), (+U0432) <Вв> is represents a voiced bilabial stop, on the other. Cyrillic letters with descenders are illustrative of glyph differences, which would be addressed in font design for an individual language.

Вв	Çç	Єє	Ьь	Зз	Зз
----	----	----	----	----	----

Table 2: Letters available in Cyrillic Range

The remaining letters (4 x 2) are a set of upper- and lower-case characters that can be enumerated in four alveolars, see Table 3. These lack the same distinctive feature, they are all simply missing a descender.

Dd	Ll	Ss	Tt
----	----	----	----

Table 3: Letters requiring descenders

To remedy future problems, perhaps a combining descender should also be added to UNICODE.

### 4. In Conclusion

Drafting a proposal for new characters such as those used in the Komi Latin alphabet (1932–1937), requires good preparation. Previous proposals should be consulted, and all problems presented should be reassessed. Make your arguments precise, and be prepared to accept assessments.

The transitional alphabet scenario is a way to allow for digital accessibility to finite though even extensive materials. For the prolific Komi language materials from the 1930s, this solution is sufficient.

It might be assumed, however, that "transitional alphabet" would be a weak argument for the Cyrillic soft sign look-alike letters <Ьь> when dealing with Ingrian (ISO-639-3 izh). This language has only been written in the Latin script.

All modifications and proposals to UNICODE come at a price, but be receptive and descriptive, and a workable solution is sure to be found.

## 5. Acknowledgements

Deborah W. Anderson is a member of the "UNICODE Script Ad Hoc" group. She has provided helpful suggestions for making a successful proposal to UNICODE Technical Committee.

Marina Fedina is the director at FU-Lab in Syktyvkar, Komi Republic, Russian Federation. She has made language materials available to us and introduced us to others involved in the digitization of Komi.

Jussi-Pekka Hakkarainen has provided us not only with access to materials from the Digitization Project of Kindred Languages but with continued advice in our work with various players.

Riitta Koikkalainen is the coordinator of the Finnish Localization Project (Kotoistushanke), and she has been instrumental in the drafting of the proposal: "Komi Latin letters missing in UNICODE".

Erkki Kolehmainen is a member of the Finnish Localization Project (Kotoistushanke), and he has provided erudite criticism and advice regarding initial drafts of proposal: "Komi Latin letters missing in UNICODE".

Enye Lav works at FU-Lab in Syktyvkar, and he has helped locate examples of various glyphs and characters from different years. He has also provided stylized representations of the missing Komi Latin letters as well as an introduction to the entire digitization process at FU-Lab.

## 6. Bibliographical References

- Grenoble, L. A. (2003). *Language policy in the Soviet Union*, volume 3. Springer Science & Business Media.
- Partanen, N. and Rießler, M. (2019). An ocr system for the unified northern alphabet. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 77–89.
- Siegel, F. and Rießler, M. (2015). Uneven steps to literacy. In Janne Saarikivi Heiko F. Marten, Michael Rießler et al., editors, *Cultural and linguistic minorities in the Russian Federation and the European Union*, pages 189–229. Springer, number 13 in Multilingual Education.
- Silfverberg, M. and Rueter, J. (2014). Can morphological analyzers improve the quality of optical character recognition? In *Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages*, pages 45–56.