

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

Doc Type: Working Group Document

Title: Archaic Komi Cyrillic characters for the BMP of the UCS

Source: Michael Everson, (EGT, IE), Klaas Ruppel (Kotus, FI), Trond Trosterud (Finsk institutt, NO)

Status: Expert Contribution

Action: For consideration by JTC1/SC2/WG2 and UTC

Date: 2000-06-09

This document requests additional characters to be added to the UCS and contains the proposal summary form.

A. Administrative

1. Title

Archaic Komi Cyrillic characters for the BMP of the UCS

2. Requester's names

Michael Everson, (EGT, IE), Klaas Ruppel (Kotus, FI), Trond Trosterud (Finsk institutt, NO)

3. Requester type

Expert contribution.

4. Submission date

2000-06-09

5. Requester's reference

6a. Completion

This is a complete proposal.

6b. More information to be provided?

No.

B. Technical -- General

1a. New script? Name?

No.

1b. Addition of characters to existing block? Name?

Yes. To the Cyrillic block or possibly a new Extended Cyrillic block.

2. Number of characters

16

3. Proposed category

Category A though they are not used in current (post-1940) Komi orthography.

4. Proposed level of implementation and rationale

Level 1; no combining characters are used.

5a. Character names included in proposal?

Yes.

5b. Character names in accordance with guidelines?

Yes.

5c. Character shapes reviewable?

Yes (see below).

6a. Who will provide computerized font?

Michael Everson, EGT.

6b. Font currently available?

Yes.

6c. Font format?

TrueType.

7a. Are references (to other character sets, dictionaries, descriptive texts, etc.) provided?

Yes, see bibliography below.

7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?

Yes, see below.

8. Does the proposal address other aspects of character data processing?

Yes, see Unicode properties below.

C. Technical -- Justification

1. Contact with the user community?

Yes. The Library community (Library of Congress, ISO TC46/SC4); Kotimaisten kielten tutkimuskeskus, Helsinki; Finsk Institut, Universitetet i Tromsø.

2. Information on the user community?

Specialists in Uralics and the Komi language, libraries.

3a. The context of use for the proposed characters?

Used to write the Komi language in the former Soviet Union, between 1919 and *ca.* 1940.

3b. Reference

See bibliography.

4a. Proposed characters in current use?

They are no longer part of current orthographies, but they are certainly found in literature and are used in library records and linguistic citations.

4b. Where?

Documents printed during the period in question.

5a. Characters should be encoded entirely in BMP?

Yes.

5b. Rationale

All Cyrillic characters should be encoded in the BMP.

6. Should characters be kept in a continuous range?

No.

7a. Can the characters be considered a presentation form of an existing character or character sequence?

No. Erroneous unifications do exist: (cf. the *ALA-LC romanization tables*).

7b. Where?

7c. Reference

8a. Can any of the characters be considered to be similar (in appearance or function) to an existing character?

No. Superficial resemblances which caused earlier unifications have been shown not to reflect actual usage and identity of these characters.

8b. Where?

8c. Reference

9a. Combining characters or use of composite sequences included?

No.

9b. List of composite sequences and their corresponding glyph images provided?

No.

10. Characters with any special properties such as control function, etc. included?

No.

E. Proposal

Komi Cyrillic characters are missing from the UCS. Document N1744 proposes the addition of 55 Cyrillic letters from ISO 10754:1996 which are not present in ISO/IEC 10646. Of those, 12 have since been added to the UCS. This proposal deals with 16 of the remaining 43 letters, which 16 were used in Komi Cyrillic orthography from 1919–ca. 1940. These letters use glyphs which differ structurally from other characters in the UCS that represent similar sounds, namely Serbian *љ* and *њ*, which are ligatures of base letters *л* and *н* with a palatalizing front-*yer* *ь*. Unification of *љ* and *њ* with *љ* and *њ* would imply that the other Komi letters *дь*, *зь*, *сь*, and *тв* could also be represented with *ь*-based glyphs, which is not supported by Komi texts using this orthography. The palatalization hook used in the Molodcov orthography is unrelated to the front *yer*.

Unification of *љ* with *љ* and *њ* with *њ* has been made in the *ALA-LC romanization tables* where the Serbian letters are given in parentheses following the Komi letters. The evidence shows this to be an error. The *ALA-LC romanization tables* also transliterate Serbian *љ* and *њ* as *lj* and *nj*, but Komi *љ* and *њ* as *ĺ* and *ń*.

Research clearly indicates that the number of Cyrillic characters in the UCS will continue to be augmented. The Current Cyrillic block does not contain enough unused positions to accept the 16 characters proposed here; accordingly, we suggest the extension of the Cyrillic block to include the next three free columns: U+0500–U+052F

Unicode Character Properties

Spacing letters, category “Lu” (uppercase), bidi category “L” (strong left to right)

0500, 0502, 0504, 0506, 0508, 050A, 050C, 050E

Spacing letters, category “Ll” (lowercase), bidi category “L” (strong left to right)

0501, 0503, 0505, 0507, 0509, 050B, 050D, 050F

Bibliography

- Barry, Randall K. 1997. *ALA-LC romanization tables: transliteration schemes for non-Roman scripts*. Washington, DC: Library of Congress Cataloging Distribution Service. ISBN 0-8444-0940-5
- Гиляревский, Р. С., & В. С. Гривнин. 1964. *Определитель языков мира по письменностям*. Москва: Наука.
- Микушев, А. К., ed. 1979. *История коми литературы*. 3 томов. Сыктывкар: Коми книжное издательство.
- Шахов, Н. А. 1924. *Краткий коми-русский словарь. С приложением статьи А. С. Сидорова „Морфологическая структура коми языка“*. Устььсыольск: Коми Издательство.

TABLE XXX - Row 05: CYRILLIC EXTENDED B

	050	051	052
0	Ԁ		
1	ԁ		
2	Ԃ		
3	ԃ		
4	Ԅ		
5	ԅ		
6	Ԇ		
7	ԇ		
8	Ԉ		
9	ԉ		
A	Ԑ		
B	ԑ		
C	Ԓ		
D	ԓ		
E	Ԕ		
F	ԕ		

G = 00
P = 00

Annex

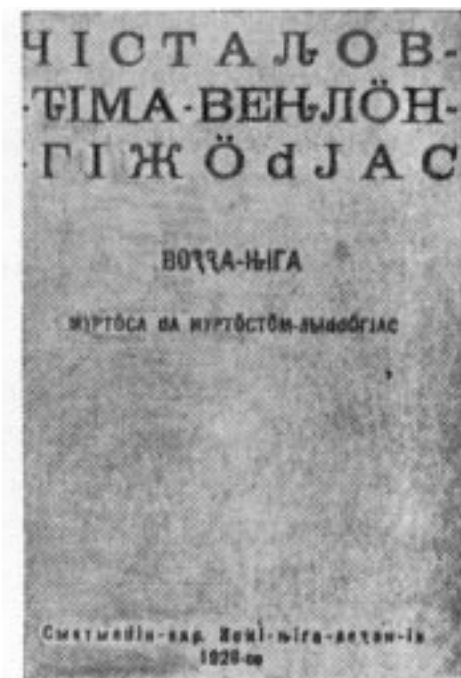


Figure 1. Cover of *Муртöса да муртöстöм лыддöгяс: вoддзa-нъгa*, by Тима Венъ (V. Čistalov) (Тима Венъ (V. Čistalov)), 1928 (modern orthography *Муртöса да муртöстöм лыддьянъяс: воддза-нъга*, by Тима Венъ (V. Čistalëv) (Тима Ven' (V. Čistalëv)). (Микушев 1979)



Figure 3. Cover of *„Чурка нина“: вит-торја драма*, by Жугыль (Žugyl') (Žugyl'), 1928 (modern orthography *„Чурка нина“: вит-торја драма*, by Жугыль (Žugyl')). Note the specifically Latin-shaped lowercase *d* in the subtitle. Note also the extremely reduced but connected form of the palatalization glyph in the stylized font in the main title – but note also that it does not have the form one would expect if it were a front yer; in which case it should show a broad square glyph like the bowl of the *P* (cf. *Ь* and Serbian *Њ* with Komi *Нь*). The transliteration into modern orthography is as given in Микушев 1979 which gives *нина* not *ньина* (presumably because of the palatalization inherent in the letter *и*).



Figure 4. Cover of *Кык повесть*, by Пеѡ Генъ (Ped' Gen'), 1936 (modern orthography *Кык повесть*, by Педь Генъ (Ped' Gen')). Note the reduced non-front-yer form of the palatalization mark in Ѣ, ѣ, Ѣ, and ѣ. Note also the shape of capital letter Ѣ as in the other examples in this annex. Modern orthography writes *повесть* rather than *повесѣть*, transferring the palatalization to the entire consonant cluster as opposed to marking each of the consonants explicitly (as in Roman orthographic practice). (Микушев 1979)



Figure 5. Cover of *Доменьлөн мыж*, by Изюр Иван (Izjur Ivan), 1936 (modern orthography *Доменьлөн мыж*, by Изьюр Иван (Iz''jur Ivan)). Note the extremely reduced size of the non-yer-like palatalization stroke in ѣ and compare it with the shape of the yer in ѣ. (Микушев 1979)



Figure 5. Cover of *Комі фольклор: важ мојдкывјас да ғыланкывјас*, 1938 (modern orthography *Коми фольклор: важ мойдъяс да съыланкывъяс*). Note the very specific form of the *л*, which is clearly unrelated to the shape of the front yer *ь*: it is not the Serbian *љ* (compare it with the yer in *ы* in *мојдкывјас*). Note also the special shape of the capital *Ѡ*; this alternates with small *d*. Despite the resemblance of Latin *d* and Komi Cyrillic *d*, the capital can certainly not be unified with Latin *D*, especially because of its relation to the palatalized pair *ѡ* and *d*. (With regard to the transliteration of *мојдкывјас* into *мойдъяс* in modern orthography, it is as given in Микυшев 1979 and is possibly a question of lexical choice, if it is not an error for *мойдкывяс*.)

Figure 6. Komi orthography as used in 1964. Additional characters for Molodcov’s orthography are also identified, with the text “To the end of the 1930s the supplementary letters *d, ѡ, ѣ, л, њ, с, т, ж, з* were used.” The text also notes that the Komi language had 230,000 speakers in 1964. (Гиляревский & Гривнин 1964:36)

А л ф а в и т

Аа, Бб, Вв, Гг, Дд, Ее, Ёё, Жж, Зз, Ии, І, Йй, Кк, Лл, Мм, Нн, Оо, Об, Пп, Рр, Сс, Тт, Уу, Фф, Хх, Цц, Чч, Шш, Щщ, ъ, Ыы, ь, Ээ, Юю, Яя.

В алфавите имеются дополнительные буквы **і, ѡ.**

Характерны окончания **-ыс, -ис, -іс.**

До конца 30-х годов употреблялись дополнительные буквы

ѡ, ѡ, ѣ, л, њ, с, т, ж, з.

Кomi (зырянский) язык распространен в Кomi АССР; на нем говорит около 230 тыс. человек. Принадлежит к финно-угорским языкам.

Figure 7. (Below and following three pages) From *Краткий коми-русский словарь*, 1924. Molodcov’s orthography, introduced into Komi schools in 1919, is used here. On page iii, the alphabetic repertoire of this orthography is given (retyped with our translation following):

Для коми языка т. Молодцов установил следующие 33 буквы: а, б, в, г, d, d̄, е, ж, ж̄, з, з̄, з̆, і, j, к, л, л̄, м, н, н̄, о, ö, п, р, с, с̄, т, т̄, у, ч, ш, щ, ы. Основую для коми шрифта послужил русский шрифт с привнесением 3-х букв (d, i, j), из латинского шрифта. Для мягких согласных выработаны свои особые обозначения: d̄ = д̄ь, з̄ = з̄ь, л̄ = л̄ь, н̄ = н̄ь, с̄ = с̄ь, т̄ = т̄ь. Буква е произносится как русское э, щ – как тш (аффрикат). Для обозначения особых звуков, не имеющих в русском языке, введены буквы ж̄, з̆, ö. Первые два из них аффрикаты, которые приблизительно можно обозначить русскими буквами следующим образом: ж̄ = дж; з̆ = дзь. При произнесении ö рот уже, чем при произнесении русского э.

For the Komi language Comrade Molodcov established the following 33 letters: а, б, в, г, d, d̄, е, ж, ж̄, з, з̄, з̆, і, j, к, л, л̄, м, н, н̄, о, ö, п, р, с, с̄, т, т̄, у, ч, ш, щ, ы. Basically the Komi script followed the Russian script with the introduction of 3 letters (d, i, j) from the Latin script. For soft consonants Komi-specific symbols were created: d̄ = д̄ь, з̄ = з̄ь, л̄ = л̄ь, н̄ = н̄ь, с̄ = с̄ь, т̄ = т̄ь. The letter е is pronounced like э, and щ like тш (affricate). For the designation of special sounds, not found in Russian, the letters were introduced ж̄, з̆, ö. The first two of these are affricates, which can be represented approximately by Russian letters in the following manner: ж̄ = дж; з̆ = дзь. For the pronunciation of ö the mouth is something like the pronunciation of Russian э.

Дополнение.

<p>Б. бажук <i>с.м.</i>—ласкательное слово. бас—украшение; басітны — украсить, украшать; басітчны—наряжаться, щеголять; басітчю, басук—щеголь, щеголиха. брунган—навозный жук.</p> <p>В. вев—фон. вежеръеш <i>неч.</i>—интерес. веслун—будничный день. вшкыны—хныкать в тихомолку. војкыа—северное сияние. ворч (пож ворч)—ободок. вбл̄d, вбл̄dja—свеже выпавший снег (пороша). вбрач—юркий, подвижный. вужас—переправа через реку.</p> <p>Г. гуза—парный комунибудь. гулыд—гладкий, ровный. гурјув, гурјів <i>в.в.</i>—открытый до отказа.</p> <p>d. діныш—комель. дружка—шафер.</p> <p>d̄. д̄авбл—дьявол.</p>	<p>д̄ад—дядя. д̄ак—псаломщик. д̄акон—дьякон. д̄ед—домовой; д̄ед лычкіс—домовой ду- шит (кошмар). д̄енга, д̄ом—деньги, монета. д̄ерт, герт—конечно. д̄етіна—мальчик. д̄иван—диван. д̄івја—не удивительно, с удовольствием. д̄івö—диво. д̄об—оставшийся последний при игре. д̄ог̄öd—деготь; д̄ог̄öд̄авны—намазать дегтем.</p> <p>Е. ежа—дерн, новина (пашня).</p> <p>Ж. жырнік—светильня из пропитанного салом фитиля.</p> <p>З. запан—запоть. зуркыд—тряска. зурьд—крепкий, не поддающийся. зыбуч—трясина.</p> <p>з̆. з̆оргыны, з̆оргіні <i>в.в.</i>—смотреть при- стально.</p>
--	--

Here we see samples from the 1924 dictionary where entries are given in a lower-case Latinski font, and the letter headings in an upper-case Bastion font. Note the distinct capital d̄ (not Д) and the distinctly non-ye shapes of the palatalization modifier in d̄, l̄, n̄, s̄, and t̄.

Ј.

јамас—осадок, речной ил (остающийся от половодья).
 јонтны—надсадить; јонтѳм—отдача от удара.
 јѳрыш—об'ем.

К.

керас—расчистка под покос или пашню.
 кесјѳдлѳм—назидание, приказывание.
 клонгысны—стукнуться твердым о твердое.
 козјан—приспособление на ремне для привешивания топора.
 којбеѳ—охотничье копье.
 крапкыны—сильно стукнуть.

Л.

лапкыштны—вспрыгнуть вниз.
 лача—надежда; сы-вылѳ лачаѳн—надеясь на него.
 лѳнѳѳ—защищенное от ветра место.

Љ.

љанѳѳ (пу)—сыроватое дерево.
 ѳуѳгыны—идти вереницей без перерыва.

М.

мајышмыны, мајѳтѳні *в.в.*—маяться.
 муцкысны—оседать от тяжести.

Н.

нымѳн—едва, чуть-чуть.

Њ.

њылѳм—пот; ѳылѳѳны—пропотеть.

О.

оль—болотистый лес на берегу реки.
 омра—вид растения из семейства зонтичных.

Ѳ.

ѳзын—подочная пристань.

П.

павкѳѳны—убеждать.
 палак *в.м.*—льдина.
 папуркі—лепешка на пресном.
 парсавны—высокомерно придираться к комунибудь.
 петас—всходы.
 пік—безвыходное положение; пікѳ воны, матѳ воны *с.*, пікје вѳні *в.в.*—попасть в безвыходное положение.
 піла пила; пілытны—пилить; пілытчыг—пильщик.
 пімі—длинные сапоги до бедер, сшитые из шкуры оленьих лапок; пімі *в.с.*—катанки.
 піѳ—зуб, зазубрина; піѳјај, піѳан—десны; піѳ јѳрны—скрежетать зубами; піѳѳѳм—беззубый.
 піѳа, агас—борона; піѳовтны, піѳајтны *печ., с.*, агсавны *н.в.*, піѳавны *н.*—боронить.
 піѳасны, відчыны *с.*—браниться; піѳавны—бранить; піѳасѳм—ссора, ругань.
 піѳовтчыны—скривиться.
 піпу—осина.
 піпѳл *л.*—см. чельаѳ.
 піс—часть саней, служащая для установки основной поверхности саней; піст *в.в.*—спица колеса, также в санях.
 піс—междометие, выражающее презрение.
 піскі, кокач *н.в.*, пістѳ—оспа; піскіавны, кокавны *н.в.*, пістѳавны—прививать оспу.
 піскѳѳны—проколоть, проложить (дорогу); пістны—проколется, найти дорогу.
 пішцалъ—ружье.
 піщѳѳ, піщег, пі *в.с.*—пазуха.
 плѳтчері, франѳч *н.*—угорь.
 плѳш, кымѳс *н.в.*—лоб.
 повны—бояться, страшиться, пугаться, трусить; повѳѳм—бесстрашный; по-

пѡч, баб—бабушка.	
празник—праздник; празничajtны—праздновать.	
прок л.—см. вын.	
прѡстѡ, вес н.—даром.	
прѡцент—процент.	
пу—дерево.	
пув—брусника; туріпув, турімомл н.—клюква; понпув—толокнянка.	
пуд—жребий; пуджавны, пуджасны—бросать жребий.	
пуд—пуд.	
пудлас—куст.	
пудовѡа, кадуля н.—четверик.	
пуж—заморожок, иней; пужжавны—заморозить.	
пужны—засучить; пужгыны—засучиться.	
пузыны, пізыны н.—вскипеть; пузѡдны, пізѡтны н.—вскипятить.	
пукны, пукавыны—сидеть; воіпукѡм—посиденка, вечеринка; пукалыс—сидящий, седок; пукгыны—сесть; пуксѡдны—посадить.	
пуктас—овоць.	
пуктыны—поставить, положить, садить; дон пуктыны—назначить цену; лыдѡ пуктыны—ставить в число, в счет, почитать; ынѡмѡ пуктыны—ставить ни во что, презирать.	
пуктысны—работать на сенокосе, садить овощи н.в.; пуктысѡм—сенокос.	
пуля—пуля.	
пуны—варить; пуѡм-пѡжалѡм—стряпня; пуыны—свариться.	
пылыыны <i>вым.</i> —быстро (торопясь) говорить.	
Р.	
распод—подсека, палыник.	
С.	
снаст—снасть.	
став—родовая группа.	
ставѡн—все вместе; ставсѡ—все; ставыс—все, все.	
сускіна—кедровый лес.	
	Г.
	гінны—одолеть; гінгыны—бороться.
	гінгыны—гнуться.
	гѡрт—крупный лес на долинах рек.
	гѡрын—с собой; гѡрас бѡстіс—взял с собой.
	гурѡс—основа.
	Т.
	тагѡс <i>вым.</i> —порог.
	тѡкмар—молот из обрубка дерева с сучком (рукояткой).
	тѡкѡтѡ—едва, чуть-чуть.
	тувкыд—упругий.
	тыѡд—запруда, лиман.
	Ѓ.
	тувкытны—быстро проскользнуть.
	У.
	уѡ—верхний снап в суслоне.
	уѡ—мягкий, сыроватый.
	уѡдны—заранее успеть.
	Ч.
	чаркі—башмачек.
	Ш.
	шаѡ!—довольно, достаточно!
	швучкан тѡв—резкий ветер.
	шупкысны—падать на мягкое.
	шуштѡм—жутко.
	шывіна—
	шыыкытны—пройти (уйти) не сказавши ни слова.
	шыпурт—меч.
	шытѡв—голос.
	Щ.
	щапкыны—схватить, зажать; щапкысны—ухватиться.
	щук—едва.