ISO/IEC JTC1/SC2/WG2 N2224
2000-06-09

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 Institutt, 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 *ӆ* and *ӊ* with *љ* and *њ* would imply that the other Komi letters *ԁ*, *ԝ*, *ԍ*, and *ҭ* 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 *ӆ* with *љ* and *ӊ* 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 *ӆ* and *ӊ* 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 статы А. С. Сидорова „Морфологическая структура коми языка“*. Устьсысольск: 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	Ԋ		
B	ԋ		
C	Ԍ		
D	ԍ		
E	Ԏ		
F	ԏ		

G = 00
P = 00

TABLE XXX - Row 05: CYRILLIC EXTENDED B

dec	hex	Name
	00	CYRILLIC CAPITAL LETTER KOMI DE
	01	CYRILLIC SMALL LETTER KOMI DE
	02	CYRILLIC CAPITAL LETTER KOMI DJE
	03	CYRILLIC SMALL LETTER KOMI DJE
	04	CYRILLIC CAPITAL LETTER KOMI ZJE
	05	CYRILLIC SMALL LETTER KOMI ZJE
	06	CYRILLIC CAPITAL LETTER KOMI DZJE
	07	CYRILLIC SMALL LETTER KOMI DZJE
	08	CYRILLIC CAPITAL LETTER KOMI LJE
	09	CYRILLIC SMALL LETTER KOMI LJE
	0A	CYRILLIC CAPITAL LETTER KOMI NJE
	0B	CYRILLIC SMALL LETTER KOMI NJE
	0C	CYRILLIC CAPITAL LETTER KOMI SJE
	0D	CYRILLIC SMALL LETTER KOMI SJE
	0E	CYRILLIC CAPITAL LETTER KOMI TJE
	0F	CYRILLIC SMALL LETTER KOMI TJE
	10	(This position shall not be used)
	11	(This position shall not be used)
	12	(This position shall not be used)
	13	(This position shall not be used)
	14	(This position shall not be used)
	15	(This position shall not be used)
	16	(This position shall not be used)
	17	(This position shall not be used)
	18	(This position shall not be used)
	19	(This position shall not be used)
	1A	(This position shall not be used)
	1B	(This position shall not be used)
	1C	(This position shall not be used)
	1D	(This position shall not be used)
	1E	(This position shall not be used)
	1F	(This position shall not be used)
	20	(This position shall not be used)
	21	(This position shall not be used)
	22	(This position shall not be used)
	23	(This position shall not be used)
	24	(This position shall not be used)
	25	(This position shall not be used)
	26	(This position shall not be used)
	27	(This position shall not be used)
	28	(This position shall not be used)
	29	(This position shall not be used)
	2A	(This position shall not be used)
	2B	(This position shall not be used)
	2C	(This position shall not be used)
	2D	(This position shall not be used)
	2E	(This position shall not be used)
	2F	(This position shall not be used)

dec	hex	Name

Annex

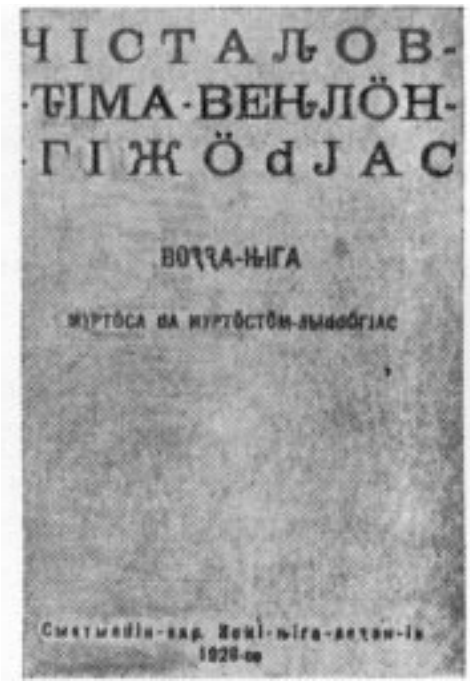

Figure 1. Cover of *Муртӧса да муртӧстӧм лыддӧгjас: воҷҷа-ҥіга,* by Ҭима Веҥ (В. Чистальов) (Ťima Veń (V. Čistaĺov)), 1928 (modern orthography *Муртӧса да муртӧстӧм лыддянъяс: воддза-ньига,* by Тима Вень (В. Чисталёв) (Tima Ven’ (V. Čistalëv)). (Микушев 1979)

Figure 3. Cover of *„Чурка ҥина“: віт-торjа драма*, by Жугыль (Žugyĺ), 1928 (modern orthography *„Чурка ньина“: вит-торйа 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 *Ҥ*). 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 повесъ*, by Педь Гень (Ped́ Geń), 1936 (modern orthography *Кык повесть*, by Педь Гень (Ped' Gen')). Note the reduced non-front-yer form of the palatalization mark in ԁ, Ԋ, G, and Ꚍ. Note also the shape of capital letter ԁ as in the other examples in this annex. Modern orthography writes *повесть* rather than *повесьть*, 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 *ԁоменьлӧн мыж*, by Ізјур Іван (Izjur Ivan), 1936 (modern orthography *Доменьлӧн мыж*, by Изъюр Иван (Iz"jur Ivan)). Note the extremely reduced size of the non-yer-like palatalization stroke in *ԋ* and compare it with the shape of the yer in *ы*. (Микушев 1979)

Figure 5. Cover of *Комі фольклор: важ мойдкывјас да сыланкывјас*, 1938 (modern orthography *Коми фольклор: важ мойдъяс да сьыланкывъяс*). Note the very specific form of the *ль*, 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 *мойдкывјас*). Note also the special shape of the capital *Ԁ*; this alternates with small *d*. Despite the resemblance of Latin *d* and Komi Cyrillic *д*, the capital can certainly not be unified with Latin *D*, especially because of its relation to the palatalized pair *Ԁь* and *ԁь*. (With regard to the transliteration of *мойдкывјас* into *мойдъяс* in modern orthography, it is as given in Микушев 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, ԁь, ҙ, ӆь, ԋь, ɡ, ҭ, ӝ, ҙ were used." The text also notes that the Komi language had 230,000 speakers in 1964. (Гиляревский & Гривнин 1964:36)

Алфавит

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

В алфавите имеются дополнительные буквы і, ӧ.

Характерны окончания -ыс, -ис, -іс.

До конца 30-х годов употреблялись дополнительные буквы ԁ, ԁь, ҙ, ӆь, ԋь, ɡ, ҭ, ӝ, ҙ.

Коми (зырянский) язык распространен в Коми АССР; на нем говорит около 230 тыс. человек.

Принадлежит к финно-угорским языкам.

86

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, ԁ, е, ж, ӂ, з, ҙ, ӡ, і, ј, к, л, ԯ, м, н, ԋ, о, ӧ, п, р, с, ҫ, т, ҭ, у, ч, ш, щ, ы. Основою для коми шрифта послужил русский шрифт с привнесением 3-х букв (d, і, ј), из латинского шрифта. Для мягких согласных выработаны свои особые обозначения: ԁ = дь, ҙ = зь, ԯ = ль, ԋ = нь, ҫ = сь, ҭ = ть. Буква е произносится как русское э, щ – как тш (африкат). Для обозначения особых звуков, не имеющихся русском языке, введены буквы ӂ, ӡ, ӧ. Первые два из них африкаты, которые приблизительно можно обозначить русскими буквами следующим образом: ӂ = дж; ӡ = дзь. При произнесении ӧ рот уже, чем при произнесении русского э.

For the Komi language Comrade Molodcov established the following 33 letters: а, б, в, г, d, ԁ, е, ж, ӂ, з, ҙ, ӡ, і, ј, к, л, ԯ, м, н, ԋ, о, ӧ, п, р, с, ҫ, т, ҭ, 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: ԁ = дь, ҙ = зь, ԯ = ль, ԋ = нь, ҫ = сь, ҭ = ть. 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 ӂ, ӡ, ӧ. The first two of these are affricates, which can be represented approximately by Russian letters in the following manner: ӂ = дж; ӡ = дзь. For the pronunciation of ӧ the mouth is something like the pronunciation of Russian э.

Дополнение.

Б.

бажук *вым.*—ласкательное слово.
бас—украшение; басітны—украсить, украшать; басітчыны—наряжаться, щеголять; басітчыс, басук—щеголь, щеголиха.
брунган—навозный жук.

В.

вев—фон.
вежертеш *печ.*—интерес.
веслун—будничный день.
вішкыны—хныкать в тихомолку.
вojкыа—северное сияние.
ворч (пож ворч)—ободок.
вӧлԁ, вӧлԁja—свеже выпавший снег (пороша).
вӧрач—юркий, подвижный.
вуӂас—переправа через реку.

Г.

гуза—парный кому нибудь.
гулыд—гладкий, ровный.
гурjув, гурjів *в.в.*—открытый до отказу.

ԁ.

ԁіныш—комель.
ԁружка—шафер.

ԁь.

ԁьавӧл—дьявол.
ԁьаԁь—дядя.
ԁьак—псаломщик.
ԁьакӧн—дьякон.
ԁьed—домовой; ԁьed лӧчкіс—домовой душит (кошмар).
ԁьеӈга, сӧм—деньги, монета.
ԁьерт, герт—конечно.
ԁьеҙіна—мальчик.
ԁьіван—диван.
ԁьівja—не удивительно, с удовольствием.
ԁьівӧ—диво.
ԁьӧб—оставшийся последний при игре.
ԁьӧгӧԁь—деготь; ԁьӧгӧԁьавны—намазать дегтем.

Е.

ежа—дерн, новина (пашня).

Ж.

жырԋік—светильня из пропитанного салом фитиля.

З.

запаԋ—запонь.
зуркыд—тряский.
зурыд—крепкий, не поддающийся.
зыбуч—трясина.

Ӡ.

ӡоргыны, ӡӧргіні *в.в.*—смотреть пристально.

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 ԁ (not Д) and the distinctly non-yer shapes of the palatalization modifier in ԁь, Ԯь, Ԋь, Ҫ, and Ҭ.

Ј.

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

К.

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

Л.

лапкыштны—вспрыгнуть вниз.
лача—надежда; сы-вылӧ лачаӧн—надеясь на него.
лӧњӧd—защищенное от ветра место.

Љ.

љанӧd (пу)—сыроватое дерево.
љуҷгыны—идти вереницей без перерыва.

М.

мајышмыны, мајітчіні *в.в.*—маяться.
мущкысны—оседать от тяжести.

Н.

нымӧн—едва, чуть-чуть.

Њ.

њылӧм—пот; њылӧdны—пропотеть.

О.

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

Ӧ.

ӧзын—лодочная пристань.

П.

павкӧdны—убеждать.
палак *вым.*—льдина.
папуркі—лепешка на пресном.
парсавны—высокомерно придираться к кому нибудь.
петас—всходы.
пік—безвыходное положение; пікӧ воны, матӧ воны *с.*, пікје воні *в.в.*—попасть в безвыходное положение.
пі́ла пила; пілі́тны—пилить; пілі́тчыс—пильщик.
пімі—длинные сапоги до бедер, сшитые из шкуры оленьих лапок; пімі *в.с.*—катанки.
піњ—зуб, зазубрина; піњјај, піњан—десны; піњ јірны—скрежетать зубами; піњтӧм—беззубый.
піња, агас—борона; піњовтны, піњајтны *печ.*, *с.*, агсавны *н.в.*, піњавны *п.*—боронить.
піњасны, відчыны *с.*—браниться; піњавны—бранить; піњасӧм—ссора, ругань.
піњовтчыны—скривиться.
піпу—осина.
піпӧл *л.*—см. челађ.
піс—часть саней, служащая для установки основной поверхности саней; пісѣт *в.в.*—спица колеса, также в санях.
піс—междометие, выражающее презрение.
піскі, кокач *н.в.*, пісѣі—оспа; піскіавны, кокавны *н.в.*, пісѣіавны—прививать оспу.
піскӧdны—проколоть, проложить (дорогу); пісѣтны—проколоться, найти дорогу.
пішщаљ—ружье.
піщӧг, піщег, пі *в.с.*—пазуха.
пљеҷчері, dрањеч *п.*—угорь.
пљеш, кымӧс *н.в.*—лоб.
повны—бояться, страшиться, пугаться, трусить; повтӧм—бесстрашный; по-

пӧч, баб—бабушка.
праӆӈік—праздник; праӆӈічајтны—праздновать.
прок *л.*—см. вын.
прӧстӧ, вес *п.*—даром.
прӧчент—процент.
пу—дерево.
пув—брусника; туріпув, турімоль *п.*—клюква; понпув—толокнянка.
пуd—жребий; пudјавны, пudјасны—бросать жребий.
пуd—пуд.
пudлас—куст.
пudовӈа, кадуӆа *п.*—четверик.
пуж—заморозок, иней; пужјавны—заморозить.
пуӂны—засучить; пуӂсыны—засучиться.
пуӆыны, піӆны *п.*—вскипеть; пуӆӧdны, піӆӧтны *п.*—вскипятить.
пукны, пукавны—сидеть; војпукӧм—посиденка, вечеринка; пукалыс—сидящий, седок; пуксыны—сесть; пуксӧdны—посадить.
пуктас—овощь.
пуктыны—поставить, положить, садить; dоп пуктыны—назначить цену; лыdӧ пуктыны—ставить в число, в счет, почитать; ӈінӧмӧ пуктыны—ставить ни во что, презирать.
пуктысны—работать на сенокосе, садить овощи *н.в.*; пуктысӧм—сенокос.
пуӆа—пуля.
пуны—варить; пуӧм-пӧжалӧм—стряпня; пусыны—свариться.
пыӆӆыны *вым.*—быстро (торопясь) говорить.

Р.

распоd—подсека, пальник.

С.

снаӧт—снасть.
став—родовая группа.
ставӧн—все вместе; ставсӧ—все; ставыс—все, все.
сускіна—кедровый лес.

Ѕ.

Ѕінны—одолеть; Ѕінсыны—бороться.
Ѕінсыны—гнуться.
Ѕӧрт—крупный лес на долинах рек.
Ѕӧрын—с собой; Ѕӧрас бостіс—взял с собой.
Ѕуröс—основа.

Т.

тагӧс *вым.*—порог.
тӧкмар—молот из обрубка дерева с сучком (рукояткой).
тӧкӧъӧ—едва, чуть-чуть.
тувкыd—упругий.
тыӧd—запруда, лиман.

Ъ.

ъувкӈітны—быстро проскользнуть.

У.

уӭ—верхний сноп в суслоне.
уӌ—мягкий, сыроватый.
уӌӧdны—заранее успеть.

Ч.

чаркі—башмачек.

Ш.

шаӈ!—довольно, достаточно!
швучкан тӧв—резкий ветер.
шупкысны—падать на мягкое.
шуштӧм—жутко.
шывіна—
шыӈкӈітны—пройти (уйти) не сказавши ни слова.
шыпурт—меч.
шытӧв—голос.

Щ.

щапкыны—схватить, зажать; щапкысны—ухватиться.
щук—едва.