| | |
|---|---|
| Title: | Comments on proposals to add characters from ISO standards developed by ISO/TC 46/SC 4 |
| Source: | NCITS/L2 and the Unicode Consortium |
| Status: | Joint Contribution |
| Action: | For the consideration of WG2 |

# 1. Introduction

### 1.1.    Overview

This contribution addresses the proposed addition to ISO/IEC 10646:1 of characters from standards for bibliographic data interchange developed by ISO/TC 46/SC 4. The characters are proposed in the following documents:

N 1741: Additional Latin characters for the UCS
N 1743: Additional Greek characters for the UCS
N 1744: Additional Cyrillic characters for the UCS
N 1745: Additional Math characters for the UCS
N 1746: Additional combining characters for the UCS
N 1747: Contraction characters for the UCS
N 1748: Additional signature mark characters for the UCS
N 1749: Additional Hebrew cantillation characters for the UCS

This contribution also comments on the proposed addition to ISO/IEC 10646 of three Cyrillic script signs that are documented in the *ALA/LC Romanization Tables* published by the Library of Congress.

### 1.2.    Background

ISO/TC 46/SC 4/WG1 wanted to ensure that all characters in the ISO/TC 46 bibliographic character sets were represented in ISO/IEC 10646, and so undertook a mapping project.  These proposals cover the characters for which WG1 was unable to identify an equivalent in ISO/IEC 10646.

The outcome of the mapping process determined the content of these proposals.  Where a mapping was overlooked, a TC 46 character was proposed for addition when it should not have been.  Questionable mappings, on the other hand, meant that a TC 46 character was not proposed for addition when perhaps it should have been.  Both these cases are addressed in comments on specific proposals.

TC 46 character sets, in general, have not been influenced by the Character/Glyph Model. Indeed, for some of the TC 46 standards, development began around the time that the distinction between character and glyph was being defined.  As a result, TC 46 standards include characters which would be considered glyphic variants under the Character/Glyph Model, and so inappropriate for encoding in ISO/IEC 10646.

The Unicode Working Group (predecessor to the Unicode Technical Committee) examined all available TC 46 standards and some DIS versions during the initial development of the Unicode

Standard.  Some TC 46 characters were not incorporated, because the Unicode Standard (like ISO/IEC 10646) conforms to the Character/Glyph Model.

Since the ISO/TC 46 character sets are used in bibliographic records, reference is made to US cataloging rules where appropriate.  Cataloging rules published by the Library of Congress and/or the American Library Association are almost universally used by American libraries.  Many libraries outside the United States also follow US practice, because of the availability of cataloging records from the Library of Congress and other US sources.

# 2. Comments on Specific Proposals

## 2.1.　N 1741: Additional Latin characters for the UCS

0220 LATIN CAPITAL LETTER KRA

No proof of the existence of an uppercase form of *kra* as an independent character in actual use has been presented.

0221 LATIN CAPITAL LETTER YR

This character is already in ISO/IEC 10646:1 as LATIN LETTER YR, encoded at position U+01A6.

LATIN SMALL LETTER YR (encoded at position 79 in ISO 5426-2:1996) has been misidentified as U+01A6 LATIN LETTER YR.

Proof of the existence of a lowercase form of *yr* as an independent character in actual use should be provided if the lowercase form is to be proposed for addition to ISO/IEC 10646.

NCITS/L2 and the Unicode Consortium note that the British Library *Special* character set which was a primary source for ISO 5426-2:1996 is consistent with the current repertoire of ISO/IEC 10646.  The British Library set (shown in Figure 1) includes only an x-height GREENLANDIC K (code E3) and  an h-height SWASH R (code EF), but not the inverse equivalents with respect to height.

## 2.2.　N 1743: Additional Greek characters for the UCS

03DB GREEK SMALL LETTER STIGMA
03DD GREEK SMALL LETTER DIGAMMA
03DF GREEK SMALL LETTER KOPPA
03E1 GREEK SMALL LETTER SAMPI

The character repertoire of Version 1.0 *of The Unicode Standard* included these characters. Examples of use have been included in N 1743 for all except *small letter sampi.*

These characters should be added to ISO/IEC 10646 with the recommended code values and names.

### 2.3. N 1744: Additional Cyrillic characters for the UCS

*Numeric Signs*

0487 CYRILLIC TEN THOUSANDS SIGN
0488 CYRILLIC HUNDRED THOUSANDS SIGN
0489 CYRILLIC MILLIONS SIGN

NCITS/L2 and the Unicode Consortium recommend:
- Unification of CYRILLIC TEN THOUSANDS SIGN with U+20DD, COMBINING ENCLOSING CIRCLE;
- Acceptance of the remaining two characters with the recommended names and code positions;

Note that all these characters are enclosing.

Comment on validity of these Cyrillic signs:
These signs are documented in the table for the transliteration of Church Slavic in the *ALA/LC Romanization Tables* (Library of Congress, 1991), p. 40-42. The ALA/LC tables were developed jointly by the Library of Congress and the American Library Association, and are used by libraries world-wide. Individual tables are evaluated by language experts during development, as well as afterwards, through actual use.

*Cyrillic Characters for non-Slavic Languages*

04C5 CYRILLIC CAPITAL LETTER CHECHEN KA
through
0519 CYRILLIC SMALL LETTER KOMI TJE.

NCITS/L2 and the Unicode Consortium oppose addition of the characters in the following ranges to ISO/IEC 10646 because they duplicate existing characters: 04C5-04C6, 04C9-04CA, 04CD-04CE, 04EC-04ED, 04F6-04F7, 04FA-04FD, 0508-0515. For details, see Appendix.

The Unicode Technical Committee and NCITS/L2 consider that the remaining letters (04FE-04FF, 0500-0507, 0516-0519) are doubtful and should not be added to ISO/IEC 10646 at this time. Further study is needed to determine whether they are truly unique or are glyphic variants of existing letters. The first pair (04FE-04FF) are from the Missionary orthography of the Mordvin-Moksha Dialect and could be the case forms of a typographic digraph; the others are all from the 1919 (Molodtsov) orthography for the Komi language.

**Background comments**

Many of the characters from ISO 10754 proposed for addition to ISO/IEC 10646 are glyphic variants of existing characters. It should be borne in mind that the initial development of ISO 10754 predates the development of the character/glyph model. ISO 10754 encodes what is *seen*, not the underlying conceptual character.

The case to add a variant form as a distinctive character in ISO/IEC 10646 must be supported by detailed, printed evidence that there is actual, contrastive use of the different forms. The case must also be made as to how Cyrillic script differs from other scripts where unification across languages applies, e.g., *Nastaliq* forms are not encoded for Urdu written in Arabic script, and language-specific forms of ideographs are encoded only because of the Source Separation Rule.

The *ALA/LC Romanization Tables* display the exact form preferred for a particular language (i.e., what is seen) because this is essential for romanization (where a Latin script string is substituted for the original Cyrillic character). The preferred letter form for a language must be shown so that

the cataloger can determine the proper Latin script substitute. The fact that the *ALA/LC Romanization Tables* display various language-specific letter forms should have no bearing on the content of a Cyrillic character repertoire formulated according to the Character/Glyph Model.

### 2.4. N 1745: Additional Math characters for the UCS

22F2 VECTOR OR SUM
22F3 VECTOR PRODUCT
22F4 SUM OR UNION OF CLASSES OR SETS
22F5 PRODUCT OF INTERSECTION OF CLASSES OR SETS
22F6 IS INCLUDED IN SET
22F7 INCLUDES IN SET

These characters are from Table 2, *Extension of Basic Set G0* of ISO 6862:1996. ISO DIS 6862 was a source for Version 1.0 of *The Unicode Standard.* In the opinion of the Unicode Working Group (predecessor to the Unicode Technical Committee), these characters replicated characters in Table 1, *Basic Set G0.* The Unicode Technical Committee re-examined the characters and came to the same conclusion. NCITS/L2 concurs.

NCITS/L2 and the Unicode Consortium oppose addition of this whole collection to ISO/IEC 10646 because they duplicate existing characters.

### 2.5. N 1746: Additional combining characters for the UCS

*Right and Left Descenders*

0346 COMBINING RIGHT DESCENDER
0347 COMBINING LEFT DESCENDER

These combining marks are intended for use with Cyrillic letters (as described in Clauses 6.2 and 6.4 of ISO 10754). Table A.2 in ISO 10754 documents combinations of Cyrillic letters and descenders. The ISO 10754 sequence of a combining descender followed by a letter can be mapped to extant characters in ISO/IEC 10646:1, except for one case pair which is proposed for addition in N 1744. The combining descenders are therefore redundant.

NCITS/L2 and the Unicode Consortium also oppose addition of these characters to ISO/IEC 10646, because of the serious implications their addition would have for decomposition, which:
- will impact the development and implementation of other standards (such as the International String Ordering standard, 14651); and,
- can destabilize the effort underway to define standard normalization forms of Unicode/10646 for use by W3C and by the programming languages community.

*Combining Small Letters Above*

0348 COMBINING LATIN SMALL LETTER A ABOVE
0349 COMBINING LATIN SMALL LETTER E ABOVE
034A COMBINING LATIN SMALL LETTER R ABOVE
034B COMBINING LATIN SMALL LETTER Z ABOVE

The only documentation for these characters is ISO 5426-2:1996 itself.

The proposed character COMBINING LATIN SMALL LETTER E ABOVE is an early form of the *umlaut.* US cataloging practice is to substitute the *umlaut* (*Descriptive Cataloging of Rare Books*, p. 69).

NCITS/L2 and the Unicode Consortium oppose addition of a character that is essentially a glyphic variant of the *umlaut*.

The remaining letters appear to be Latin contractions (see comments on N 1747 below).


*Multiple Diacritical Marks*

034C COMBINING DOUBLE CARON
034D COMBINING DOUBLE CIRCUMFLEX
034E COMBINING CIRCUMFLEX GRAVE

The only documentation for these characters is ISO 5426-2:1996 itself.

These marks appear to be Latin contractions (see comments on N 1747 below).


### 2.6.     N 1747: Contraction characters for the UCS

00 LATIN CONTRACTION AGUS (et, ond)
through
15 LATIN SMALL CONTRACTION LONG S WITH HOOK

also, 2048 REVERSED SECTION SIGN in N 1748, and possibly many of the combining characters in N 1746.

*LATIN CONTRACTION AGUS*

Mr. Michael Everson has said that this character is the modern Irish equivalent of the ampersand. The British Library's "Special" character set from which ISO 5426-2 is partially derived calls this character "Ampersand form type 2." *Descriptive Cataloging of Rare Books* (2nd. ed., Library of Congress, 1991) instructs the cataloger (p. 69):

> If the Tironian sign (<image of Tironian sign>) cannot be reproduced, treat it as an abbreviation and substitute "[et]" for it.

This existence of this character is documented by three sources.  The US rules for rare book cataloging prescribe its use.  It should be added to ISO/IEC 10646, but as a General Punctuation character with the proposed name TIRONIAN SIGN.

*Mappable Latin Contractions*

06 LATIN CONTRACTION REVERSED US
07 LATIN CONTRACTION IS
08 LATIN CONTRACTION SMALL IS
09 LATIN CONTRACTION UM

These characters can be mapped to existing UCS characters and so should be eliminated from the proposal.

| Prop. No. | ISO 5426-2 | | ISO/IEC 10646:1 | |
|---|---|---|---|---|
| | Code | Name | Code | Name |
| 09 | 02/14 | CONTRACTION MARK LATIN CAPITAL LETTER SCRIPT I | U+2110 | SCRIPT CAPITAL I |
| 08 | 02/15 | CONTRACTION MARK HEAVY APOSTROPHE | U+02BC | MODIFIER LETTER APOSTROPHE |
| 07 | 03/12 | CONTRACTION MARK LATIN SMALL LETTER SCRIPT OPEN E | U+025C | LATIN SMALL LETTER REVERSED OPEN E |
| 06 | 03/13 | CONTRACTION MARK LATIN SMALL LETTER SCRIPT E | U+212F | SCRIPT SMALL E |

*General Comments about Latin Contractions*

US practice for rare book cataloging is to substitute the spelled-out equivalent for the contraction. Rule 0J2 of *Descriptive Cataloging of Rare Books* (p. 7) states:

> When special marks of contraction have been used by the printer in continuance of the manuscript tradition, expand affected words to their full form and enclose supplied letters in square brackets. When an abbreviation standing for an entire word appears in the source, record instead the word itself, and enclose it in square brackets.

(The exception to this rule, noted above, is the Tironian sign.)

The US position is that these contraction signs should not be added to ISO/IEC 10646. Contraction signs of this general sort might be useful additions as glyphic forms for fonts used in the facsimile reproduction of manuscripts and early printed books, but their encoding as characters for the representation of textual content is counterproductive. Their use for the representation of text would cause searching and comparison problems in electronic texts and cataloging records so encoded.

Furthermore, the Latin contractions in ISO 5426-2 are only a small subset of the contractions used over the centuries in manuscripts and later included in some early printed works. Contractions are not exclusive to Latin manuscripts. Examples of Cyrillic contractions appear in the documentation accompanying N 1744. There is no good rationale to merely pick out the small subset of such manuscript contraction forms present in ISO 5426-2 and propose them for encoding, as opposed to any other set which could be brought forward.

While the Unicode Consortium and NCITS/L2 do not recommend the encoding of **any** of these contraction signs (other than the *agus*/Tironian sign) in ISO/IEC 10646, if WG2 chooses to accept any of the remaining characters proposed in N 1747 (i.e., 01-15 and 0A-15), then they should:
a)  be considered Letterlike Symbols, for use in any script context; and,
b)  be given descriptive names (e.g., those in ISO 5426-2:1996) rather than names based on the meaning of the contraction sign in a particular language context.
The names proposed in N 1747 are generally unacceptable.

### 2.7. N 1748: Additional signature mark characters for the UCS

2048 REVERSED SECTION SIGN
2049 REVERSED PILCROW SIGN
2139 LATIN CAPITAL LETTER ROTATED Q
2183 ROMAN NUMERAL REVERSED ONE HUNDRED
2614 SIX-SPOKED ASTERISK
2615 BLACK LEFTWARDS BULLET
2616 BLACK RIGHTWARDS BULLET
2768 REVERSED ROTATED FLORAL HEART BULLET

Editorial: The first two cells in the code chart extract are misnumbered.

*Problematic Characters*

2048 REVERSED SECTION SIGN

This character is a Latin contraction, not a signature mark.

The image of this character is incorrect (as is its name, which is based on the incorrect image). ISO 5426-2:1996 says that this is 'Used for the Latin suffix "orum".' The British Library's "Special" character set includes a character "-RUM WORD ENDING TYPE 2" which may be the origin of this character.

If characters from ISO 5426-2 representing Latin contractions are added to ISO/IEC 10646:1, this character should not be added until it has been properly identified.

2614 SIX-SPOKED ASTERISK

CHASE is an EU-funded project to develop Unicode/UCS mappings for character sets used in European libraries. CHASE has mapped the British Library's equivalent of this character to U+2736, SIX POINTED BLACK STAR. This character should be eliminated from the proposal.

*Acceptable Characters*

The following characters should be added to ISO/IEC 10646:1, each with the proposed name and code value:

2049 REVERSED PILCROW SIGN
2139 LATIN CAPITAL LETTER ROTATED Q
> The image of this character should more closely resemble to the British Library source character.
2183 ROMAN NUMERAL REVERSED ONE HUNDRED
> This character is not a section symbol but a Roman numeral. Mr. Michael Everson provided convincing examples of use.

The following characters should be added to ISO/IEC 10646:1, with the proposed names but in the General Punctuation block.

2615 BLACK LEFTWARDS BULLET
2616 BLACK RIGHTWARDS BULLET
2768 REVERSED ROTATED FLORAL HEART BULLET

General Comments about Signature Marks

The marks used by printers to identify signatures (also called collations or gatherings) are important for the study of early printing.  The repertoire of signature marks in ISO 5426-2 is only a small subset of the marks used by printers.  Since ISO/IEC 10646 is intended to encode plain text, and the particular font used in a signature mark may be significant, the full and correct representation of signatures marks cannot be achieved without the use of a higher level protocol. The Unicode Consortium and NCITS/L2 note that the bullets and the reversed pilcrow sign are general-purpose punctuation marks.


### 2.8.    N 1749: Additional Hebrew cantillation characters for the UCS

05F5 HEBREW ACCENT TSERE
through
05FC HEBREW ACCENT ASTERISK

These proposed characters are from Table 2: *Special vowel points, accents, and marks* of ISO 8957:1996.  The contents of Table 2 are historic cantillation marks.  There are no known implementations of this table.

The decision to equate certain Hebrew cantillation marks with diacritical marks (for example, 41 HEBREW ACCENT TSERE was mapped to U+0308 COMBINING DIAERESIS)
is questionable

WG1 also mapped 5A HEBREW ACCENT RAFE to U+05BF HEBREW POINT RAFE, even though the two characters had different names and images.  Furthermore, 4C HEBREW POINT RAFE in Table 1 is mapped to U+05BF HEBREW POINT RAFE, so that the net effect is to unify 4C in the *Basic Hebrew alphabet* and 5A in the *Special vowel points, accents, and marks*.

NCITS/L2 and the Unicode Consortium recommend that the addition of characters from Table 2 of ISO 8957:1996 be deferred.  This will allow experts on cantillation marks and historic pointing and scholars from Israel and from other countries to clarify the nature of specific marks, and to determine the correct mappings for certain TC 46 characters.  This will allow a complete proposal for any missing characters to be developed.

**Background information on library implementations of Hebraic script**

Except for the name of one character, Table 1: *Basic Hebrew alphabet* corresponds to the USMARC Hebrew character set published by the Library of Congress. It is used by US libraries with significant collections of Hebraica.

This character set was developed by the Research Libraries Group and subsequently adopted by the Library of Congress for USMARC.  To determine the repertoire of characters, RLG consulted librarians responsible for Hebraica collections.  The librarians unanimously rejected the inclusion of cantillation marks.

A European implementation of Table 1 is under development for the Consortium of European Research Libraries.

In Israel, library data is encoded using an SII character set similar to ISO 8859-8, rather than ISO 8957.  Nevertheless, the Standards Institution of Israel did participate in the development of ISO 8957.

# 3. References:

ISO 5426-2:1996, *Information and documentation – Extension of the Latin alphabet coded character set for bibliographic information interchange – Part 2: Latin characters used in minor European languages and obsolete typography.*

ISO 6862:1996, *Information and documentation – Mathematical coded character set for bibliographic information interchange.*

ISO 10754:1996, *Information and documentation – Extension of the Cyrillic alphabet coded character set for non-Slavic languages for bibliographic information interchange*

*ALA/LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts.* Tables compiled and edited by Randall K. Barry. Library of Congress, 1997.

*Descriptive Cataloging of Rare Books,* 2nd. ed., Library of Congress, 1991.

R. S. Gilyarevsky & V. S. Grivnin, *Languages Identification Guide.* Nauka, 1970.

# Appendix: Characters from ISO 10754:1996

Disapprove addition for the reason shown in bold face:

04C5   CYRILLIC  CAPITAL LETTER CHECHEN KA
04C6   CYRILLIC SMALL LETTER CHECHEN KA
**Already encoded**.  The uppercase *Chechen ka* is U+041A + U+030A and the lowercase
*Chechen ka*  is U+043A + U+030A.
Editorial: The images at 04C5 and 04C6 in N 1744 are incorrect; the appendage is a circle, not a curve.

04C9   CYRILLIC CAPITAL LETTER CHUVASH NG
04CA   CYRILLIC SMALL LETTER CHUVASH NG
**Already encoded**.  The *Chuvash ng* is a glyphic variant of either of these Cyrillic letters: *nje* or *en with descender.*  (*ALA/LC Romanization Tables*, p. 118)

04CD   CYRILLIC CAPITAL LETTER KOMI NG
04CE   CYRILLIC SMALL LETTER KOMI NG
**Already encoded**.  The *Komi ng* is a glyphic variant of  the Cyrillic letter *nje.*  (ISO 10754:1996, Table A.1, also *ALA/LC Romanization Tables*, p. 123)

Annex B of ISO 10754:1996 states: In some cases, early efforts to "cyrillicize" the writing of certain languages made use of some letters from other alphabets."  The following four characters exemplify this.

04EC   CYRILLIC CAPITAL LETTER SELKUP OE
04ED   CYRILLIC SMALL LETTER SELKUP OE
**Already encoded.**

04F6   CYRILLIC CAPITAL LETTER AISOR EL
04F7   CYRILLIC SMALL LETTER AISOR EL
**Already encoded.**

04FA   CYRILLIC CAPITAL LETTER KURDISH QA
04FB   CYRILLIC SMALL LETTER KURDISH QA

**Already encoded.**

04FC  CYRILLIC CAPITAL LETTER KURDISH WE
04FD  CYRILLIC SMALL LETTER KURDISH WE
**Already encoded.**

0508  CYRILLIC CAPITAL LETTER YAKUT I WITH STROKE
0509  CYRILLIC SMALL LETTER YAKUT I WITH STROKE
**Already encoded**.  The uppercase *Yakut I with stroke* is U+0406 + U+0335  and the lowercase *Yakut I with stroke* is U+0456 + U+0335.

050A  CYRILLIC CAPITAL LETTER JE WITH STROKE
050B  CYRILLIC SMALL LETTER JE WITH STROKE
**Already encoded**.  The uppercase *Je with stroke* is U+0408 + U+0335  and the lowercase *Je with stroke* is U+0458 + U+0335.

050C  CYRILLIC CAPITAL LETTER KOMI ELJ
050D  CYRILLIC SMALL LETTER KOMI ELJ
**Already encoded**.  The *Komi elj* is a glyphic variant of  the Cyrillic letter *lje. .*  (ISO 10754:1996, Table A.1, also *ALA/LC Romanization Tables*, p. 123)

050E  CYRILLIC CAPITAL LETTER EL WITH MIDDLE HOOK
050F  CYRILLIC SMALL LETTER EL WITH MIDDLE HOOK
**Already encoded**.  The *el with middle hook* is a glyphic variant of  the Cyrillic letter *lje.* (*ALA/LC Romanization Tables*, p. 118)

0510  CYRILLIC CAPITAL LETTER MORDVIN EL KA
0511  CYRILLIC SMALL LETTER MORDVIN EL KA
ALA/LC: Mordvin-Moksha dialect (1923) (=Lkh/lkh with dot on h)

**Typographical digraph.** "The letter combinations **k[**, **h[** are characteristic of the Mordvin-Moksha language." (Gilyarevsky & Grivnin, p. 42).  The discrete letters are already encoded.

Editorial comment: These characters should be named CYRILLIC CAPITAL LIGATURE EL HA and CYRILLIC SMALL LIGATURE EL HA respectively, based on (a) the quoted information, and (b) the ALA/LC romanizations for the case forms.  In the basic Cyrillic letters at the beginning of the ALA/LC romanization table for non-Slavic languages (p. 114), **K/k** is romanized as L/l, **{ /[** is romanized as Kh/kh, which agrees with the romanizations for the case forms of the digraph in the table for *Mordvin-Moksha Dialect (1923)* (p. 127), i.e., Lkh/lkh (with dot on h).

0512  CYRILLIC CAPITAL LETTER EN WITH MIDDLE HOOK
0513  CYRILLIC SMALL LETTER EN WITH MIDDLE HOOK
**Already encoded**.  The letter is listed for Chuvash and Khanty (Vahk) in Table A.1 of ISO 10754:1996, but is not included in ALA/LC romanization tables for these alphabets.  Treat as a glyphic variant of  the Cyrillic letter *nje.*

0514  CYRILLIC CAPITAL LETTER MORDVIN ER KA
0515  CYRILLIC SMALL LETTER MORDVIN ER KA
**Typographical digraph.** "The letter combinations **k[**, **h[** are characteristic of the Mordvin-Moksha language." (Gilyarevsky & Grivnin, p. 42).  The discrete letters are already encoded.

Editorial comment: These characters should be named CYRILLIC CAPITAL LIGATURE ER HA and CYRILLIC SMALL LIGATURE ER HA respectively, based on (a) the quoted information, and (b) the ALA/LC romanizations for the case forms.  In the basic Cyrillic letters at the beginning of the ALA/LC romanization table for non-Slavic languages (p. 114), **H/h** is romanized as R/r, **{ /[** is

romanized as Kh/kh, which agrees with the romanizations for the case forms of the digraph in the table for *Mordvin-Moksha Dialect (1923)* (p. 127), i.e., Rkh/rkh (with dot on h).


The following proposed characters require further study:

```
04FE    CYRILLIC CAPITAL LETTER YA IE
04FF    CYRILLIC SMALL LETTER YA IE
0500    CYRILLIC CAPITAL LETTER KOMI DE
0501    CYRILLIC SMALL LETTER KOMI DE
0502    CYRILLIC CAPITAL LETTER KOMI DJE
0503    CYRILLIC SMALL LETTER KOMI DJE
0504    CYRILLIC CAPITAL LETTER KOMI DZE
0505    CYRILLIC SMALL LETTER KOMI DZE
0506    CYRILLIC CAPITAL LETTER KOMI ZJE
0507    CYRILLIC SMALL LETTER KOMI ZJE
0516    CYRILLIC CAPITAL LETTER KOMI ESJ
0517    CYRILLIC SMALL LETTER KOMI ESJ
0518    CYRILLIC CAPITAL LETTER KOMI TJE
0519    CYRILLIC SMALL LETTER KOMI TJE
```