

TO: The Unicode Technical Committee and ISO/IEC JTC1/SC2 WG2
FROM: Deborah Anderson (SEI, UC Berkeley) on behalf of Tapani Salminen
DATE: 30 January 2012
RE: Request for 2 New Cyrillic Characters for the Khanty and Nenets Languages

L2/12-052

Summary:

Tapani Salminen, a Finno-Ugrian language specialist with expertise in the Forest Nenets, Northern Khanty, and Eastern Khanty languages, provides evidence in this document from recent native publications for 2 Cyrillic characters that are not yet in Unicode, but which are needed by the native communities in Siberia to represent their languages.

The two characters are:

Л̑ CYRILLIC CAPITAL LETTER EL WITH DESCENDER
л̑ CYRILLIC SMALL LETTER EL WITH DESCENDER

Background on the encoding of EL WITH DESCENDER

The CYRILLIC LETTER EL WITH DESCENDER, a fricolateral, is only found in a small number of languages in the Far North (Siberia), including Northern Khanty, Eastern Khanty, and Forest Nenets.

Because the history of encoding of this character is closely tied to encoding CYRILLIC EL characters with various “appendages”, the discussion below traces the history of the encoding of these other characters.

The current set of forms of the Cyrillic letter EL with an “appendage” as they appear in the Unicode Standard include:

04C5 Л̑ CYRILLIC CAPITAL LETTER EL WITH TAIL
04C6 л̑ CYRILLIC SMALL LETTER EL WITH TAIL
• Kildin Sami

Khanty letters

0512 ЈЈ CYRILLIC CAPITAL LETTER EL WITH HOOK
0513 јј CYRILLIC SMALL LETTER EL WITH HOOK
• also used for Chukchi and Itelmen

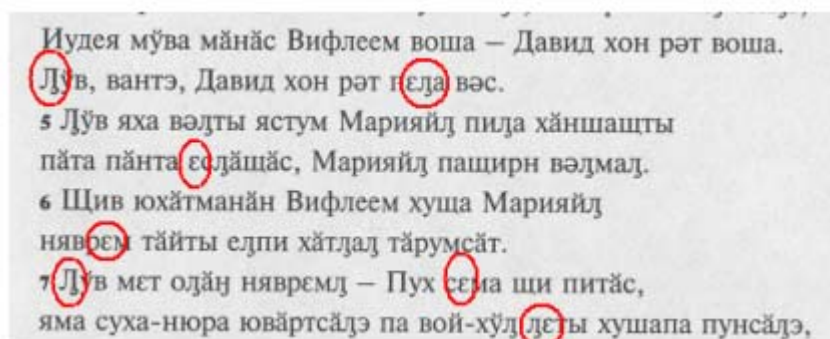
The following characters are not yet encoded:

----- Л̑ CYRILLIC CAPITAL LETTER EL WITH DESCENDER
----- л̑ CYRILLIC SMALL LETTER EL WITH DESCENDER

The upper and lowercase EL WITH DESCENDER characters were first proposed in 1997 (N1590) and then

again in 1998 (N1744¹) as a character needed for the Kildin Sámi language. The small and capital characters were approved by the character encoding committees and appeared on a ballot for an amendment to 10646 (ISO/IEC 10646-1: 1993/Amd. 30: 1999²). However, subsequent research identified the “appendage” as a hook, not a descender³, so the EL WITH DESCENDER characters were withdrawn. Small and capital EL WITH TAIL were later proposed (N2173⁴ in 2000), approved, and appeared in Unicode 3.2 at the code points 04C5 and 04C6. In short: No EL WITH DESCENDER characters were encoded. (The EL WITH DESCENDER character does not appear in Kildin Sámi.)

In 2005, a proposal (L2/05-080) for various characters for languages of the Far North was submitted (and its characters approved). The proposal included EL WITH HOOK, which was identified as a character used in Khanty. One example of the EL character with an “appendage” in running text in Khanty was provided (from page 5 of proposal, see figure below). The example came from a publication by the Institute for Bible Translation⁵, and showed an EL character with a hook-shaped glyph:



The proposal document also included a figure from an article by Berdnikov⁶, who surveyed the use of the Cyrillic script in various writing systems. The example from Berdnikov, which came from a chart of generalized data used in “some national alphabets” (Berdnikov: 40), identified the EL WITH HOOK glyph as being used in “Hanty (kazym)” [Khanty] as well as Chukcha [sic, =Chukchi] and Itelmen, but provided no running text sample.⁷

¹ <http://std.dkuug.dk/jtc1/sc2/wg2/docs/n1744.pdf>.

² For the FPDAM, see <http://www.itscj.ipsj.or.jp/sc2/open/02n3309.pdf>

³ See the 1999 document “When is a descender not a descender: Kildin Sámi voiceless consonants”, available at: <http://www.hum.uit.no/a/trond/kildinbackgr.pdf>.

⁴ See “Proposal to add 8 Cyrillic Sami characters to ISO/IEC 10646” from NTS (Norway), SFS (Finland), and NSAI (Ireland), dated 2003-03-03, available at: <http://www.unicode.org/L2/L2000/00082-n2173.pdf>.

⁵ Institute for Bible Translation. *Rozhdestvo Iisusa Khrista*. Moskva : Institut perevoda Biblii, 2000. The Institute for Bible Translation agency is based in Stockholm.

⁶ Page 6 of Berdnikov, A. Mars. 1998. *Alphabets Necessary for Various Cyrillic Writing Systems*. Cahiers GUTenberg no. 28-29 – Congrès EuroTEX.

⁷ The article by Berdnikov did note the issues in the development and use of Cyrillic for various languages, saying “further confusion arises as the glyphs used to represent some letter [sic] have been changed from time to time” and he then lists as an example the EL WITH DESCENDER glyph “<->” EL WITH HOOK glyph, which suggests variation on the shape of the “appendage”. Berdnikov goes on to say: “Finally, it is not uncommon for there to be several projected alphabets for a single language, and for different publishing houses to use different alphabets.” (Berdnikov: 38) This seems to reflect the situation for Khanty.

In the proposal (page 7), the authors distinguished the tail from the descender or hook modification, but added that the linguistic sources consulted did not consider the descender to be a variant of a hook.

In response to the proposal, a second document was submitted (L2/05-215) by D. Anderson, which provided feedback from various scholars and experts. Several scholars strongly advised that the user communities be directly consulted. The feedback document also recommended verification be sought for Khanty. Unfortunately, no follow-up was done to gather further input on Khanty or to obtain examples from the native communities until contact was made with Tapani Salminen in late 2011 on this topic.

Because the original proposal for EL WITH HOOK cited only one figure for Khanty with running text, and contained no examples published in the native territory, we recommend the UTC carefully review the discussion with examples provided below and consider encoding EL WITH DESCENDER, as well as making changes to existing annotations.

Use of EL WITH DESCENDER in the Khanty and Nenets languages

Tapani Salminen reports that in native publications of the Northern Khanty, Eastern Khanty, and Forest Nenets languages, the EL WITH DESCENDER is used (see figs. 1, 2, 3).

The appearance of the hook in some publications, particularly those outside of the native communities, was due to the predominance of a particular font (“Prosveshchenie”) used by a publisher in St. Petersburg, which in turn influenced a few other publications. In the Prosveshchenie font, the “appendage” appears as a hook. However, in common typographic practice, Northern Khanty, Eastern Khanty, Forest Nenets, and Tundra Nenets use a descender.⁸

In sum, without EL WITH DESCENDER it is impossible to compose coherent electronic publications in Northern Khanty, Eastern Khanty and Forest Nenets, among other languages. Lack of EL WITH DESCENDER also poses problems for searching. EL WITH DESCENDER has been used by members of indigenous nations for the writing of their native languages for several decades, and there is great potential for electronic publishing that would directly benefit from having the full range of the local alphabets covered by Unicode.

Use of descenders on other letters in the Khanty and Nenets languages

Although the request is specifically to encode EL WITH DESCENDER, it is to be noted that descenders appear on other letters used by the Khanty and Nenets languages (see images from native publications):

- EN WITH DESCENDER [U+04A2/3] used in Northern Khanty (fig.1), Eastern Khanty (fig. 2), Forest Nenets (fig. 3), Tundra Nenets (fig. 4)
- KA WITH DESCENDER [U+49A/B] used in Eastern Khanty (fig. 2)
- CHE WITH DESCENDER [U+04B6/7] used in Eastern Khanty (fig. 2)

⁸ An example of the influence of the Prosveshchenie font appears in figure 5, taken from a dictionary of the Forest Nenets published in Moscow, in which the EL WITH DESCENDER and EN WITH DESCENDER appear with a hook-like shape. Compare the Forest Nenets examples in figure 3, which is printed in the native territory.

Based on this evidence, the annotations for certain “hook” characters in the Unicode Standard need to be revised. “Khanty” should be removed from the following characters (as well as U+0512/3 EL WITH HOOK and possibly U+0510 and U+0511 CYRILLIC REVERSED ZE, see footnote 9):

- U+04C3/4 KA WITH HOOK
- U+0478/8 EN WITH HOOK

Character Properties

05DE;CYRILLIC CAPITAL LETTER EL WITH DESCENDER;Lu;0;L;;;;;N;;; 05DF;
05DF;CYRILLIC SMALL LETTER EL WITH DESCENDER;Ll;0;L;;;;;N;;; 05DE;;05DE

Collation

CYRILLIC LETTER EL WITH DESCENDER should sort after EL WITH TAIL (04C6) and before EL WITH HOOK (0513), following the pattern for CYRILLIC LETTER EN.

FIGURES

Figure 1. Below is an image of Northern Khanty, from the Хӧнты ясаң newspaper № 52 (3184; 27 December 2008). (Хӧнты ясаң is a newspaper published in Northern Khanty, Eastern Khanty, and Forest Nenets.) As can be seen in the figure, the current Northern Khanty orthography incorporates EL WITH DESCENDER, EN WITH DESCENDER, and UKRAINIAN IE. The characters known as EL WITH HOOK, EN WITH HOOK, and REVERSED ZE⁹ are not used in common Northern Khanty orthography. (Source: http://www.helsinki.fi/~tasalmin/Northern_Khanty.jpg)

Валентина Нико-
лаевна, тӧлаң ёш,
тӧлаң кӧр наңена!

⁹ Hence, the current annotation for U+0510 and U+0511 CYRILLIC REVERSED ZE, which says “Khanty”, should be removed, pending further study.

Figure 2. The image below comes from the same issue of the Хӑнты ясаң newspaper as figure 1, but the language on this page is Eastern Khanty. Its current orthography incorporates KA WITH DESCENDER, EL WITH DESCENDER, EN WITH DESCENDER, and CHE WITH DESCENDER. The character known as KA WITH HOOK, just like EL WITH HOOK and EN WITH HOOK discussed above, is not used in common Eastern Khanty orthography. (Source: http://www.helsinki.fi/~tasalmin/Eastern_Khanty.jpg)

Қул көнччө қө

Әй пухәлднә қул кәнччө ики вёл. Әймәта латнә қул кәнччө ики йӑвөны ай рытнат мән. Сарнә йӑвөн йӑчөнә сорәм мөх. Ванхә йөвөт, қоләп қөчөхтәта йәх. Қоләп қөв. Лӑрпиләтәх, лӑрпиләтәх, әй латнә вӑрхә ти йәх. Ар қулднә питы. Қуләт рыта пӑн, йӑңқәнам мән. Ики пухәла йөвөт, утә қил.

ӑдал. Кӑнчәт қөвөт. Вӑс икинә пырипи: «Нӑң қөяхи вөсән?» «Ма ӑяң қө вөсәм. Қӑтя, вөв йичалди!» – пырәс ньӑвмиләд. Ӓяң қө ястәд: «Вӑс икия вө-вам әнтә йөвөтл. Ма пөткахтәләм, вӑр ар тӑйләм.

«Лӑв тӑӑқа тенә! – пытәмтәх Вӑс ики. – Рытә махәлта кирәх-

«Қулпәм лӑрпитәтәхә пыхәрта! Вӑс ики қоләп лӑрпитәтәхә ымәд. Ӓяң қөнә лупнат ухәди мӑтқәмты. Вӑс ики йөңка кӑрәх. Йөмат кәчақди. Атәм турат вӑккәтәх. Өс тю пырәс ики қынт илмөхтәх панә йӑқәнам нӑрәхтәх.

Руслан
КУРЛОМКИН

Figure 3. The figure below is from still the same issue of the Хӑнты ясаң newspaper as figure 2, but the language is Forest Nenets, whose orthography has EL WITH DESCENDER and EN WITH DESCENDER. What was said above about EL WITH HOOK and EN WITH HOOK is valid in the case of Forest Nenets as well. The phrase in parentheses in Eastern Khanty for ‘in the Nenets language’, and the last three words following the author’s name are in Eastern Khanty as well.¹⁰ (Source: http://www.helsinki.fi/~tasalmin/Forest_Nenets.jpg)

Талӑха

(ЙӑРХАН ЯСӘН)

Ҙамы тӑхананта талӑха тидимай. Пытта ваңкшахана тидиңа. Чоняку тоңа, ваңкшахата питалата. Талӑха куняңә кай” не. Петалахана тятылӑ, тятылӑ. Кимяхалт ничахаст мят ңайптаман. Талӑха пӑутең танай, хӑв-

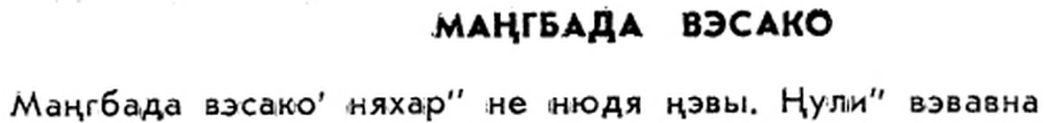
кат маныңа, маншту: пӑ”ниня ваңкшатай шелта. Куптаң хөвкат пытяханта ваңкшаутем шелтапича. Талӑха ваңкшантай пытай. Ваңкшаутета хомяхама, чекехента пытта тидина. Хөвкат мячеш тошту.

Анжела ИСЛАМОВА

Варәң йӑвөн пухәд

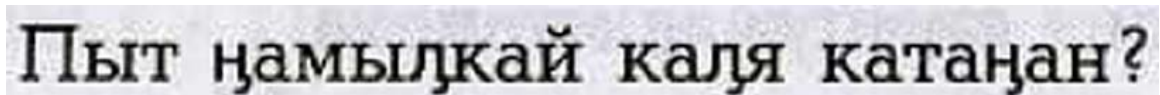
¹⁰ Note: The letter SCHWA WITH DIAERESIS is a provisional replacement of E WITH DIAERESIS, i.e. ӓ, in regular Forest Nenets orthography, based on Khanty influence.

Figure 4. The image below comes from *Фольклор народов Таймыра, выпуск 2: ненецкий фольклор* (составитель К. И. Лабанаускас; Дудинка 1992). This folklore collection in Tundra Nenets was published locally in the Tundra Nenets country, and it continues employing regular, classic typography for EN WITH DESCENDER. Other publications created in the native territory, both printed and electronic, in Tundra Nenets and in neighbouring indigenous languages, follow similar practice. (Source: http://www.helsinki.fi/~tasalmin/FN92_30.jpg)



МАҢБАДА ВЭСАКО
Маңбада вэсако' няхар'' не нюдя ңэвы. Ңули'' вэавна

Figure 5. The image below comes from *Русско-ненецкий словарь (лесной диалект): Пособие для учителей и учащихся начальных классов* (Москва: Икар, 1997) by Е. Н. Вожакова. This tiny dictionary of the Forest Nenets language, published in Moscow, shows an orthography with both EN WITH DESCENDER and EL WITH DESCENDER. The design of characters is based on the classic model but influenced by the Prosveshchenie font. Since EL WITH DESCENDER is currently not included in Unicode, local authors resort to ad hoc solutions such as using the Kildin Saami characters with a “tail” which, however, do not fit the purpose either formally or functionally. (Source: <http://www.helsinki.fi/~tasalmin/Vozhakova.jpg>)



Пыт ңамылкай каля катаңан?

**ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹**

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html> for guidelines and details before filling this form.

Please ensure you are using the latest Form from <http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest Roadmaps.

A. Administrative

1. Title: **Request for 2 New Cyrillic Characters for the Khanty and Nenets Languages**

2. Requester's name: *Deborah Anderson (SEI, UC Berkeley) on behalf of Tapani Salminen*

3. Requester type (Member body/Liaison/Individual contribution): *Liaison*

4. Submission date: *1 February 2012*

5. Requester's reference (if applicable):

6. Choose one of the following:

This is a complete proposal: YES

(or) More information will be provided later:

B. Technical – General

1. Choose one of the following:

a. This proposal is for a new script (set of characters): NO

Proposed name of script:

b. The proposal is for addition of character(s) to an existing block: YES

Name of the existing block: *Cyrillic Supplement*

2. Number of characters in proposal:

3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary B.1-Specialized (small collection) B.2-Specialized (large collection)

C-Major extinct D-Attested extinct E-Minor extinct

F-Archaic Hieroglyphic or Ideographic G-Obscure or questionable usage symbols

4. Is a repertoire including character names provided? Yes

a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? Yes

b. Are the character shapes attached in a legible form suitable for review? Yes

5. Fonts related:

a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?

--

b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):

6. References:

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

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

7. Special encoding issues:

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? Yes

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at <http://www.unicode.org> for such information on other scripts. Also see Unicode Character Database (<http://www.unicode.org/reports/tr44/>) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

¹ Form number: N4102-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? If YES explain	<i>Yes</i> <i>See discussion in proposal; characters were originally proposed in 1997 but later withdrawn</i>
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? If YES, with whom? If YES, available relevant documents:	<i>Yes</i> <i>Expert, Tapani Salminen</i>
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Reference:	<i>See below</i> <i>Ethnologue reports 13,600 speakers of Khanty and 31,300 of Nenets (Forest and Tundra)</i>
4. The context of use for the proposed characters (type of use; common or rare) Reference:	<i>common</i> <i>See proposal</i>
5. Are the proposed characters in current use by the user community? If YES, where? Reference:	<i>yes</i> <i>Newspapers, etc.</i>
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? If YES, is a rationale provided? If YES, reference:	<i>no</i>
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	<i>yes</i>
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? If YES, is a rationale for its inclusion provided? If YES, reference:	<i>No</i>
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? If YES, is a rationale for its inclusion provided? If YES, reference:	<i>No</i>
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? If YES, is a rationale for its inclusion provided? If YES, reference:	<i>See proposal</i>
11. Does the proposal include use of combining characters and/or use of composite sequences? If YES, is a rationale for such use provided? If YES, reference: Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? If YES, reference:	<i>No</i>
12. Does the proposal contain characters with any special properties such as control function or similar semantics? If YES, describe in detail (include attachment if necessary)	<i>No</i>
13. Does the proposal contain any Ideographic compatibility characters? If YES, are the equivalent corresponding unified ideographic characters identified? If YES, reference:	<i>No</i>