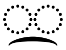
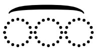
ISO/IEC JTC1/SC2/WG2 N3571
L2/09-028
2009-01-27

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

Doc Type: **Working Group Document**
Title: **Proposal to encode additional characters for the Uralic Phonetic Alphabet**
Source: **Klaas Ruppel, Tero Aalto, Michael Everson**
Status: **National Body Contribution (Finland and Ireland)**
Date: **2009-01-27**

Request. The 11 combining characters and subscript letters proposed here date from the time before the formal standardization of UPA in 1902. However, experience in digitalizing large corpora has revealed the need for extending the standard to these "pre-UPA" characters. That is to say, there is scope for the symbols listed below to be used in documents using the current UPA.

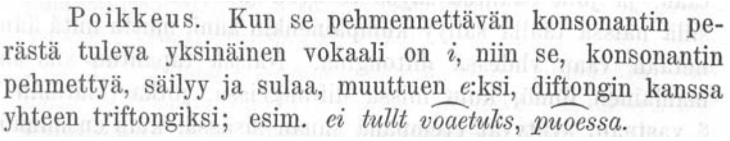
	1DFA	COMBINING DOUBLE INVERTED BREVE BELOW
	1DFB	COMBINING TRIPLE INVERTED BREVE
h	A7F2	LATIN SUBSCRIPT SMALL LETTER H
k	A7F3	LATIN SUBSCRIPT SMALL LETTER K
l	A7F4	LATIN SUBSCRIPT SMALL LETTER L
m	A7F5	LATIN SUBSCRIPT SMALL LETTER M
n	A7F6	LATIN SUBSCRIPT SMALL LETTER N
p	A7F7	LATIN SUBSCRIPT SMALL LETTER P
s	A7F8	LATIN SUBSCRIPT SMALL LETTER S
t	A7F9	LATIN SUBSCRIPT SMALL LETTER T
ꟺ	A7FA	LATIN LETTER SMALL CAPITAL TURNED M

Unicode character properties.

```
1DFA;COMBINING TRIPLE INVERTED BREVE;Mn;234;NSM;;;;;N;;;;;
1DFB;COMBINING DOUBLE INVERTED BREVE BELOW;Mn;233;NSM;;;;;N;;;;;
A7F2;LATIN SUBSCRIPT SMALL LETTER H;Ll;0;L;<sub> 0068;;;;N;;;;;
A7F3;LATIN SUBSCRIPT SMALL LETTER K;Ll;0;L;<sub> 006B;;;;N;;;;;
A7F4;LATIN SUBSCRIPT SMALL LETTER L;Ll;0;L;<sub> 006C;;;;N;;;;;
A7F5;LATIN SUBSCRIPT SMALL LETTER M;Ll;0;L;<sub> 006D;;;;N;;;;;
A7F6;LATIN SUBSCRIPT SMALL LETTER N;Ll;0;L;<sub> 006E;;;;N;;;;;
A7F7;LATIN SUBSCRIPT SMALL LETTER P;Ll;0;L;<sub> 0070;;;;N;;;;;
```

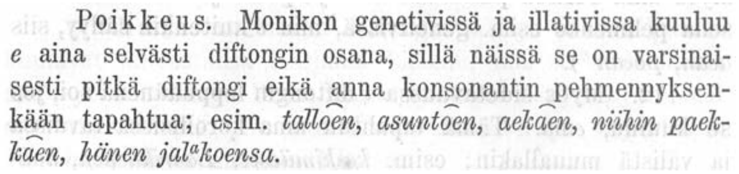
```
A7F8;LATIN SUBSCRIPT SMALL LETTER S;Ll;0;L;<sub> 0073;;;;N;;;;;
A7F9;LATIN SUBSCRIPT SMALL LETTER T;Ll;0;L;<sub> 0074;;;;N;;;;;
A7F;LATIN LETTER SMALL CAPITAL TURNED M;Ll;0;L;;;;;N;;;;;
```

Figures.



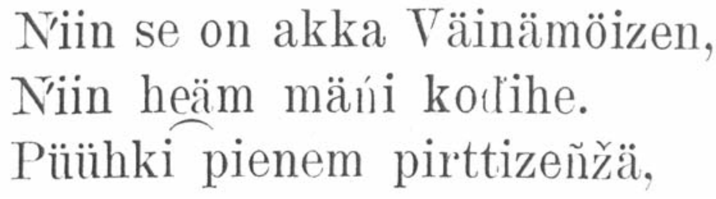
Poikkeus. Kun se pehmennettävän konsonantin perästä tuleva yksinäinen vokaali on *i*, niin se, konsonantin pehmettyä, säilyy ja sulaa, muuttuen *e*:ksi, diftongin kanssa yhteen triftongiksi; esim. *ei tullt voaetuks*, *puoessa.*

Figure 1. The diacritic COMBINING TRIPLE INVERTED BREVE indicates a triphhong (a group of three vowels in one syllable). Example from Ossian Grotenfelt, “Pohjois-Hämeen kielimurteesta”, in *Suomi* II:12 (page 321, rows 22–25). This character has also been proposed in N3555, also at U+1DFB, along with ◌◌◌ COMBINING TRIPLE BREVE BELOW at U+1DFC.



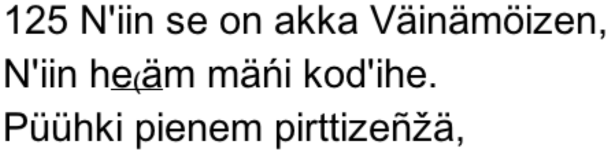
Poikkeus. Monikon genetivissä ja illativissa kuuluu *e* aina selvästi diftongin osana, sillä näissä se on varsinaisesti pitkä diftongi eikä anna konsonantin pehmennyksenkään tapahtua; esim. *tallo͡en*, *asunto͡en*, *aeka͡en*, *niihin paekka͡en*, *hänen jal*^{a}*koensa.*

Figure 2. The existing diacritic U+0361 COMBINING DOUBLE INVERTED BREVE is used in the same document to mark a diphthong. Example from Ossian Grotenfelt, “Pohjois-Hämeen kielimurteesta”, in *Suomi* II:12 (page 320, rows 3–7).



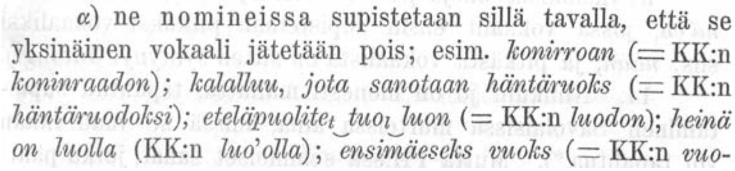
N'iin se on akka Väinämöizen,
N'iin heäm mäńi kod'ihe.
Püühki pienem pirttizeňžä,

Figure 3a. The diacritic COMBINING DOUBLE INVERTED BREVE BELOW indicates a syllable boundary between vowels. From *Suomen Kansan Vanhat Runot*. I. *Vienan läänin runot*. I. Kalevala-aineiset kertovaiset runot. Suomalaisen Kirjallisuuden Seuran toimituksia 121:1. Helsinki 1908 (page 5, verses 125–127). This book the above example originates in has been entirely digitized and is publicly available online (http://dbgw.finlit.fi/skvr/). In the electronic version, the proposed character is currently replaced by a fallback solution shown in Figure 3b:



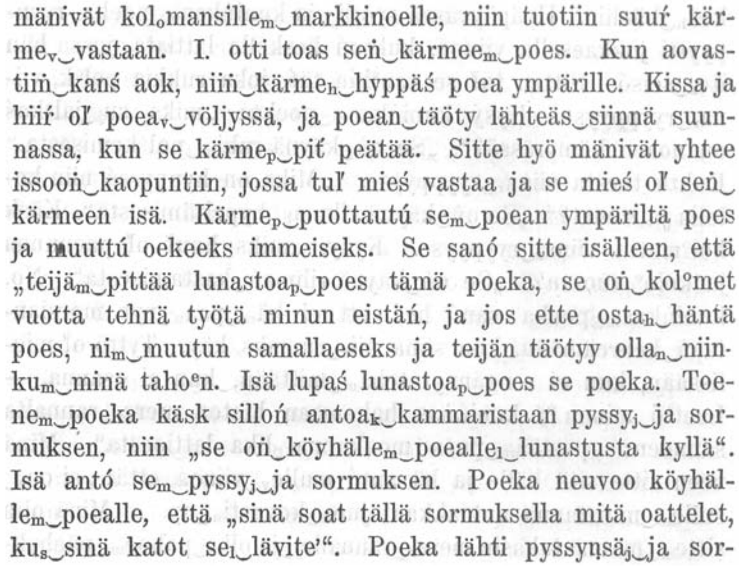
125 N'iin se on akka Väinämöizen,
N'iin heäm mäńi kod'ihe.
Püühki pienem pirttizeňžä,

Figure 3b. Fallback for COMBINING DOUBLE INVERTED BREVE BELOW, taken from http://dbgw.finlit.fi/skvr/showtext.phtml?id=skvr01100020



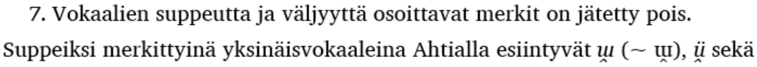
α) ne nomineissa supistetaan sillä tavalla, että se yksinäinen vokaali jätetään pois; esim. *konirroan* (= KK:n *koninraadon*); *kalalluu, jota sanotaan häntäruoks* (= KK:n *häntäruodoksi*); *eteläpuolite$_t$‿tuo$_l$‿luon* (= KK:n *luodon*); *heinä on luolla* (KK:n *luo'olla*); *ensimäeseks vuoks* (= KK:n *vuo-*

Figure 4. Example showing LATIN SUBSCRIPT SMALL LETTER T and LATIN SUBSCRIPT SMALL LETTER L, from Ossian Grotenfelt, "Pohjois-Hämeen kielimurteesta", in *Suomi* II:12 (page 321). These and other subscript characters indicate that the initial consonant of the next word modifies the previous word's end in the speech stream. The final sound of the previous word is phonetically assimilated towards the initial consonant of the following word (e.g. *sen pyssyn* → *sem pyssyn*; *niin muuten* → *niim muuten*).



mänivät kol$_o$mansille$_m$‿markkinoelle, niin tuotiin suuŕ kärme$_v$‿vastaan. I. otti toas seṅ‿kärmee$_m$‿poes. Kun aovastiiṅ‿kanś aok, niiṅ‿kärme$_h$‿hyppäś poea ympärille. Kissa ja hiiŕ oľ poea$_v$‿völjyssä, ja poean‿täöty lähteäs‿siinnä suunnassa, kun se kärme$_p$‿piť peätää. Sitte hyö mänivät yhtee issooṅ‿kaopuntiin, jossa tuľ mieś vastaa ja se mieś oľ seṅ‿kärmeen isä. Kärme$_p$‿puottautú se$_m$‿poean ympäriltä poes ja muuttú oekeeks immeiseks. Se sanó sitte isälleen, että „teijä$_m$‿pittää lunastoa$_p$‿poes tämä poeka, se oṅ‿kol^{o}met vuotta tehnä työtä minun eistäń, ja jos ette osta$_h$‿häntä poes, ni$_m$‿muutun samallaeseks ja teijän täötyy olla$_n$‿niinku$_m$‿minä taho^{e}n. Isä lupaś lunastoa$_p$‿poes se poeka. Toene$_m$‿poeka käsk silloń antoa$_k$‿kammaristaan pyssy$_j$‿ja sormuksen, niin „se oṅ‿köyhälle$_m$‿poealle$_l$‿lunastusta kyllä". Isä antó se$_m$‿pyssy$_j$‿ja sormuksen. Poeka neuvoo köyhälle$_m$‿poealle, että „sinä soat tällä sormuksella mitä oattelet, ku$_s$‿sinä katot se$_l$‿lävite'". Poeka lähti pyssyṇsä$_j$‿ja sor-

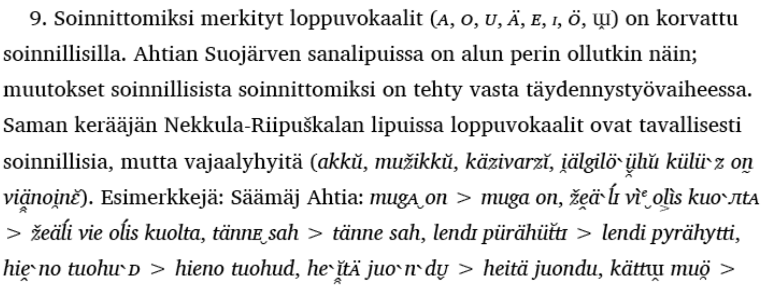
Figure 5. Example showing LATIN SUBSCRIPT SMALL LETTER H, U+2C7C LATIN SUBSCRIPT SMALL LETTER J, LATIN SUBSCRIPT SMALL LETTER K, LATIN SUBSCRIPT SMALL LETTER M, LATIN SUBSCRIPT SMALL LETTER N, LATIN SUBSCRIPT SMALL LETTER P, LATIN SUBSCRIPT SMALL LETTER S, and U+1D65 LATIN SUBSCRIPT SMALL LETTER V. From Ossian Grotenfelt, "Pohjois-Hämeen kielimurteesta", in *Suomi* II:12 (page 366).



7. Vokaalien suppeutta ja väljyyttä osoittavat merkit on jätetty pois.

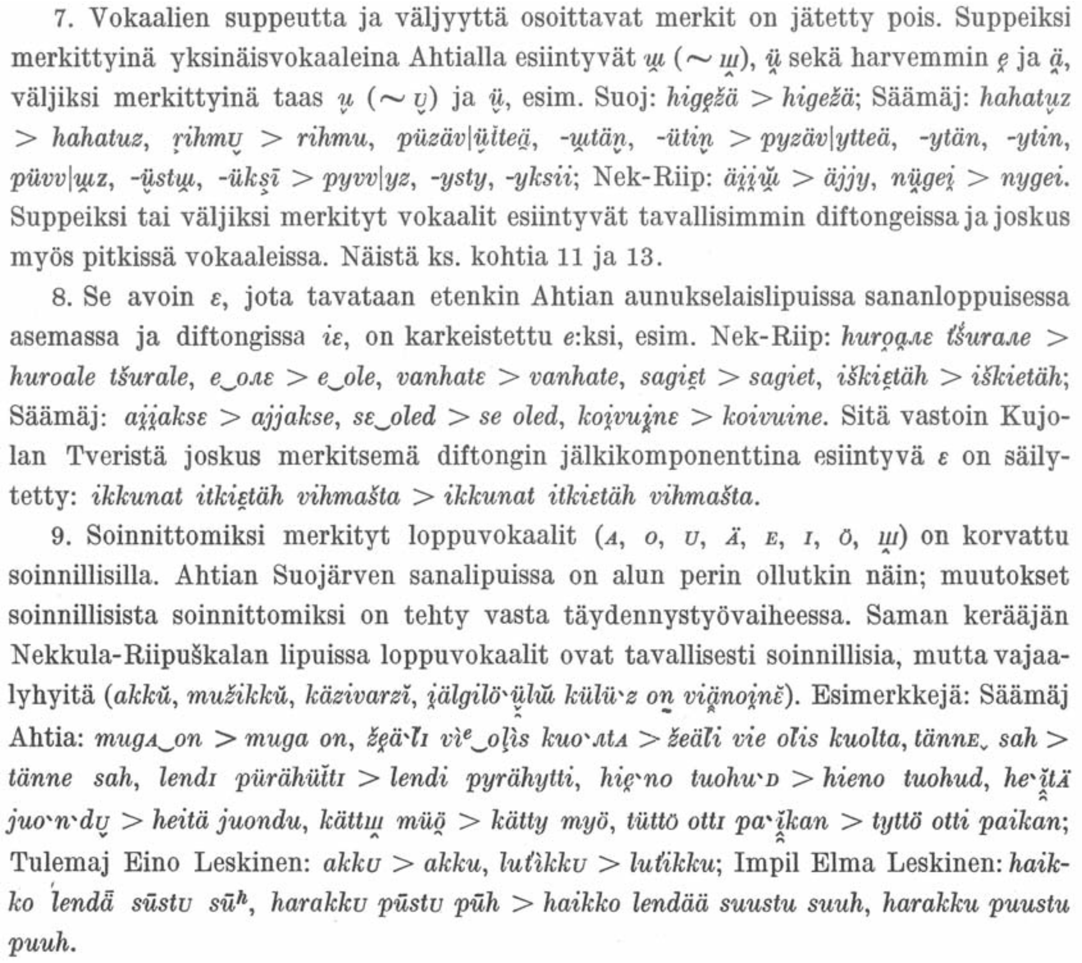
Suppeiksi merkittyinä yksinäisvokaaleina Ahtialla esiintyvät *ɯ̭* (~ ɯ̭), *ṷ̈* sekä

Figure 6a. Example a Karelian dictionary is currently being digitized for online publication. A pilot version is already available (http://kaino.kotus.fi/sanat/kks/). In the online version, the proposed character LATIN LETTER SMALL CAPITAL TURNED M has, for now, been replaced with a lower case unitalicized letter, as opposed to the predominant convention of italicizing UPA content—but this is just a fallback. From http://kaino.kotus.fi/sanat/kks/KKS_Johdanto.pdf



9. Soinnittomiksi merkityt loppuvokaalit (ᴀ, ᴏ, ᴜ, ᴀ̈, ᴇ, ɪ, ᴏ̈, ꟺ̭) on korvattu soinnillisilla. Ahtian Suojärven sanalipuissa on alun perin ollutkin näin; muutokset soinnillisista soinnittomiksi on tehty vasta täydennystyövaiheessa. Saman kerääjän Nekkula-Riipuškalan lipuissa loppuvokaalit ovat tavallisesti soinnillisia, mutta vajaalyhyitä (*akkŭ, mužikkŭ, käzivarzĭ, i̭älgilö˴ü̬lŭ külü˴z on̰ vi̭ä̂noi̭nɛ̆*). Esimerkkejä: Säämäj Ahtia: *mugᴀ‿on > muga on, žḙä˴ĺɪ vìᵉ‿oĺìs kuo˴лtᴀ > žeäĺi vie oĺis kuolta, tännᴇ‿sah > tänne sah, lendɪ pürähüt̆tɪ > lendi pyrähytti, hiḙ˴no tuohu˴ᴅ > hieno tuohud, he˴ĭ̭tᴀ̈ juo˴n˴du̬ > heitä juondu, kättꟺ̭ muö̭ >*

Figure 6b. Another example from the Karelian dictionary.



7. Vokaalien suppeutta ja väljyyttä osoittavat merkit on jätetty pois. Suppeiksi merkittyinä yksinäisvokaaleina Ahtialla esiintyvät ɯ̭ (∼ ꟺ̭), ṷ̈ sekä harvemmin ḙ ja ä̭, väljiksi merkittyinä taas u̬ (∼ ᴜ̬) ja ü̬, esim. Suoj: *higḙžä > higežä*; Säämäj: *hahatu̬z > hahatuz, ri̭hmᴜ̬ > rihmu, püzäv|ü̬lteä̬, -ɯ̭tän̬, -ütin̬ > pyzäv|ytteä, -ytän, -ytin, püvv|ɯ̭z, -ü̬stɯ̭, -üks̬ī > pyvv|yz, -ysty, -yksii*; Nek-Riip: *äi̭i̭ṷ̈ > äjjy, nṷ̈gei̭ > nygei.* Suppeiksi tai väljiksi merkityt vokaalit esiintyvät tavallisimmin diftongeissa ja joskus myös pitkissä vokaaleissa. Näistä ks. kohtia 11 ja 13.

8. Se avoin *ɛ*, jota tavataan etenkin Ahtian aunukselaislipuissa sananloppuisessa asemassa ja diftongissa *iɛ*, on karkeistettu *e*:ksi, esim. Nek-Riip: *huro̭a̭лɛ ťšuraлe > huroale tšurale, e‿oлɛ > e‿ole, vanhatɛ > vanhate, sagiḙt > sagiet, iškiḙtäh > iškietäh*; Säämäj: *ai̭i̭aksɛ > ajjakse, sɛ‿oled > se oled, koi̭vui̭nɛ > koivuine.* Sitä vastoin Kujolan Tveristä joskus merkitsemä diftongin jälkikomponenttina esiintyvä *ɛ* on säilytetty: *ikkunat itkiḙtäh vihmašta > ikkunat itkiɛtäh vihmašta.*

9. Soinnittomiksi merkityt loppuvokaalit (ᴀ, ᴏ, ᴜ, ᴀ̈, ᴇ, ɪ, ᴏ̈, ꟺ̭) on korvattu soinnillisilla. Ahtian Suojärven sanalipuissa on alun perin ollutkin näin; muutokset soinnillisista soinnittomiksi on tehty vasta täydennystyövaiheessa. Saman kerääjän Nekkula-Riipuškalan lipuissa loppuvokaalit ovat tavallisesti soinnillisia, mutta vajaalyhyitä (*akkŭ, mužikkŭ, käzivarzĭ, i̭älgilö˴ü̬lŭ külü˴z on̰ vi̭ä̂noi̭nɛ̆*). Esimerkkejä: Säämäj Ahtia: *mugᴀ‿on > muga on, žḙä˴ľɪ vìᵉ‿oľìs kuo˴лtᴀ > žeäľi vie oľis kuolta, tännᴇ̬ sah > tänne sah, lendɪ pürähüt̆tɪ > lendi pyrähytti, hiḙ˴no tuohu˴ᴅ > hieno tuohud, he˴ĭ̭tᴀ̈ juo˴n˴du̬ > heitä juondu, kättꟺ̭ müö̭ > kätty myö, tüttö ottɪ pa˴ĭ̭kan > tyttö otti paikan*; Tulemaj Eino Leskinen: *akkᴜ > akku, luťìkkᴜ > luťikku*; Impil Elma Leskinen: *haikko ˊlendǟ sūstᴜ sūʰ, harakkᴜ pūstᴜ pūh > haikko lendää suustu suuh, harakku puustu puuh.*

Figure 7. Example of LATIN LETTER SMALL CAPITAL TURNED M from the original Karelian dictionary, From *Karjalan kielen sanakirja*. 1. - Lexica Societatis Fenno-Ugricae XVI,1. Helsinki 1968 (page LXXXIII). In the Uralic Phonetic Alphabet, small capital letters denoting vowels refer to voiceless articulation. The turned m indicates a *closed central u*, whereas the small capital equivalent indicates a *voiceless closed central u*. Note the difference in glyph realization between capitals, smalls, and small capitals: M m ᴍ *M m ᴍ* (LETTER M), Ɯ ɯ ꟺ *Ɯ ɯ ꟺ* (LETTER TURNED M, really a kind of *uu* ligature).

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
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:** ***Proposal for Encoding 12 Additional Characters of the Uralic Phonetic Alphabet (UPA)***

2. Requester's name: *Klaas Ruppel, Tero Aalto, Michael Everson*

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

4. Submission date: *2009-01-27*

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):
Proposed name of script:
b. The proposal is for addition of character(s) to an existing block: *YES*
Name of the existing block: *Phonetic Extensions / Latin Extended-C*

2. Number of characters in proposal: *12*

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) *X*
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. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Michael Everson*
If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: *everson@evertype.com, www.evertype.com*

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)? *NO*

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 http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3152-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)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *NO*

If YES explain

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.)? *YES*

If YES, with whom? *Juhani Lehtiranta, Álgu Project (RILF)*

If YES, available relevant documents:

3. Information on the user community for the proposed characters (for example:
size, demographics, information technology use, or publishing use) is included? *YES*

Reference: *This proposal*

4. The context of use for the proposed characters (type of use; common or rare) *Linguistic*

Reference:

5. Are the proposed characters in current use by the user community? YES

If YES, where? Reference: *Álgu Project (RILF), Finno-Ugrian Society*

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely
in the BMP? *YES*

If YES, is a rationale provided? *YES*

If YES, reference: *The characters are a part of UPA*

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *NO*

8. Can any of the proposed characters be considered a presentation form of an existing
character or character sequence? *NO*

If YES, is a rationale for its inclusion provided?

If YES, reference:

9. Can any of the proposed characters be encoded using a composed character sequence of either
existing characters or other proposed characters? *NO*

If YES, is a rationale for its inclusion provided?

If YES, reference:

10. Can any of the proposed character(s) be considered to be similar (in appearance or function)
to an existing character? *NO*

If YES, is a rationale for its inclusion provided?

If YES, reference:

11. Does the proposal include use of combining characters and/or use of composite sequences? *YES*

If YES, is a rationale for such use provided? *YES*

If YES, reference: *This proposal*

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *YES*

If YES, reference: *This proposal*

12. Does the proposal contain characters with any special properties such as
control function or similar semantics? *NO*

If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility character(s)? *NO*

If YES, is the equivalent corresponding unified ideographic character(s) identified?

If YES, reference: