ISO/IEC JTC1/SC2/WG2 N2958

2005-07-22

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

Doc Type: **Working Group Document**
Title: **Report on progress in implementing the Uralic Phonetic Alphabet with indication of the need for additional characters and symbols**
Source: **Juhani Lehtiranta, Klaas Ruppel, Toni Suutari, Trond Trosterud**
Status: **Expert Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2005-07-22**

1. A proposal for the addition of characters used in the Uralic Phonetic Alphabet (UPA) was produced in 1999 as a result of a joined effort of experts in Uralistics. Finally in 2002 about 100 characters were added into the UCS. Immediately after this an international project "Unicode Implementation of Uralic and International Phonetic Alphabets (UIUIPA)" was established for promoting and improving the implementation of the UCS in Uralistics. The project members are Michael Everson and Juhani Lehtiranta (font providers), Kazuto Matsumura and Klaas Ruppel. In the beginning the work was financed mainly by the Department of Dynamic Linguistics of the University of Tokyo (before 2004: Department of Asian and Pacific Linguistics). Other institutions, as e.g. the Research Institute for the Languages of Finland (RILF) and the University of Helsinki, have provided manpower for the project.

So far the project can report as results and status of its work:

- Mono space font Everson Mono supports UPA.
- A UPA supporting font with advanced typography (including Regular and Italic face) is ready to be published for free download.
- A large survey of the frequency of special characters and combinations of characters and diacritical marks and/or modifier letters in texts written in UPA transcription has been carried out. The results will be used as a base for the development of a UPA keyboard layout.
- A detailed guide for writing UPA is ready for publishing (for the time being only in Finnish).

At RILF the UPA scheme was implemented immediately after the new characters were added in 2002. The most important project in this regard is an etymological data base for the Sámi Languages (Álgu). For the time being the data base holds about 60,000 entries, but the number will increase significantly in the course of future work. Later the data base will be made publicly available as a scientific archive (however, see below part 3).

As a default font Everson Mono is used. As an interface for the MySQL data base best results are so far achieved by using Safari under Mac OS X. Figure 1 shows the Skolt Sámi (koltansaame, ko) word for 'shaft' and its etymological correspondences (=) in Inari Sámi (in) and Ter Sámi (tj). Additionally the Skolt Sámi word is compared (~) to another Skolt Sámi word.

Figure 1: Screenshot from the Álgu database showing the Skolt Sámi word for 'shaft' and some related etymological data.

Figure 2: Screenshot from the Álgu database showing the Kildin Sámi word for 'shell'.

A current drawback regarding the display of characters is that multiple diacritics do not stack properly when presented in HTML (Figure 2). Compare with the correct stacking in the writing field. As far as we know this bug has unfortunately not been fixed in the recent updates of Safari (2.0) and the new Mac OS X (10.4).

The input method for characters used so far in Álgu is not satisfactory. The method is based on macros designed in the project. In order to achieve a better input method not only for the Álgu project but for the whole scientific community the UIUIPA project aimes to develop a UPA keyboard layout. For the practical work a Finnish sub working group consisting of representatives of RILF (Klaas Ruppel, Toni Suutari), the University of Helsinki (Tapani Lehtinen), the Finno-Ugrian Society (Leena Huima) and Juhani Lehtiranta was founded, and a survey of the frequency of UPA characters in texts written in Uralic, Turkish, Mongolian and other languages has been carried out. The survey was financed by the Finnish Association for Scholarly Publishing, the Finno-Ugrian Society and RILF.

As a result of the survey:

- UIUIPA will be able to publish a package containing a font and a keyboard layout accompanied with detailed guidance by the end of this year (initially in Finnish; English and German translations to be added).
- A few characters will be looked at further to determine, whether they can be dealt with otherwise or if they should be brought up for addition into the UCS. In the following a short discussion of those characters is given.

2. There are four characters for which the survey indicates a need for adding them into the UCS.

2.1 Two spacing modifier letters, which indicate the coincidence of main stress and high tone (A) and the coincidence of secondary stress and low tone (B). Compare the Spacing Modifier Letters 02F9 MODIFIER LETTER BEGIN HIGH TONE (C) and 02FB MODIFIER LETTER BEGIN LOW TONE (D) (Figure 3):

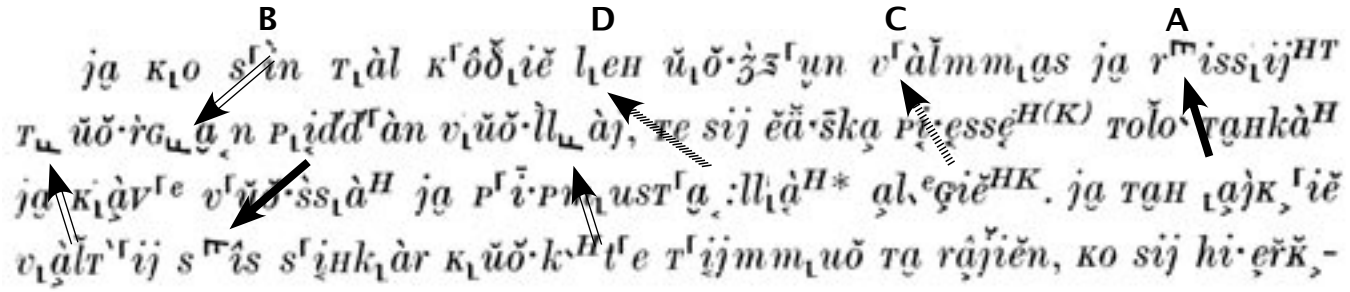


Figure 3: Lappische Volksdichtung II. Herausgegeben von Eliel Lagercrantz. — MSFOu 115. Helsinki 1958.

2.2 Two combining diacritical marks.

2.2.1 COMBINING LEFT ARROWHEAD ABOVE is in use at least in Estonian dialects. The diacritic marks the longest stage of the 3 stages of vowel stage graduation (Figure 4).

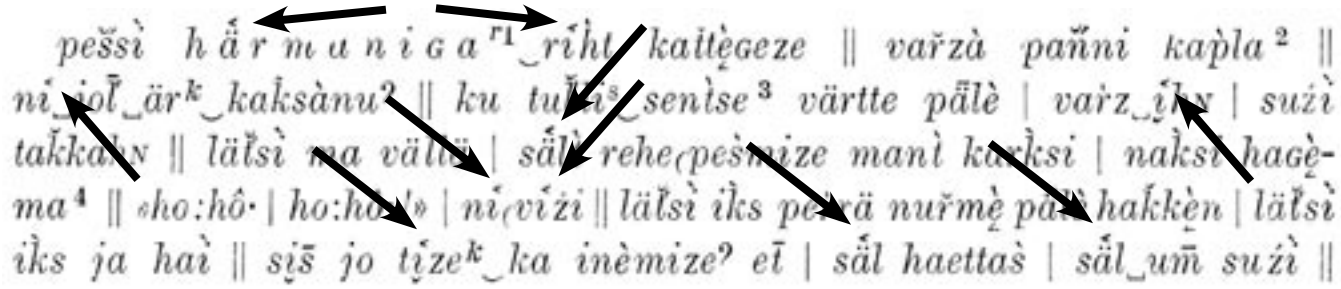


Figure 4: Pertti Virtaranta: Lähisukukielten lukemisto. — Suomalaisen Kirjallisuuden Seuran toimituksia 287. Helsinki 1967.

2.2.2 COMBINING RIGHT ARROWHEAD AND DOWN ARROWHEAD BELOW. Figure 5 gives an example from the Komi language.

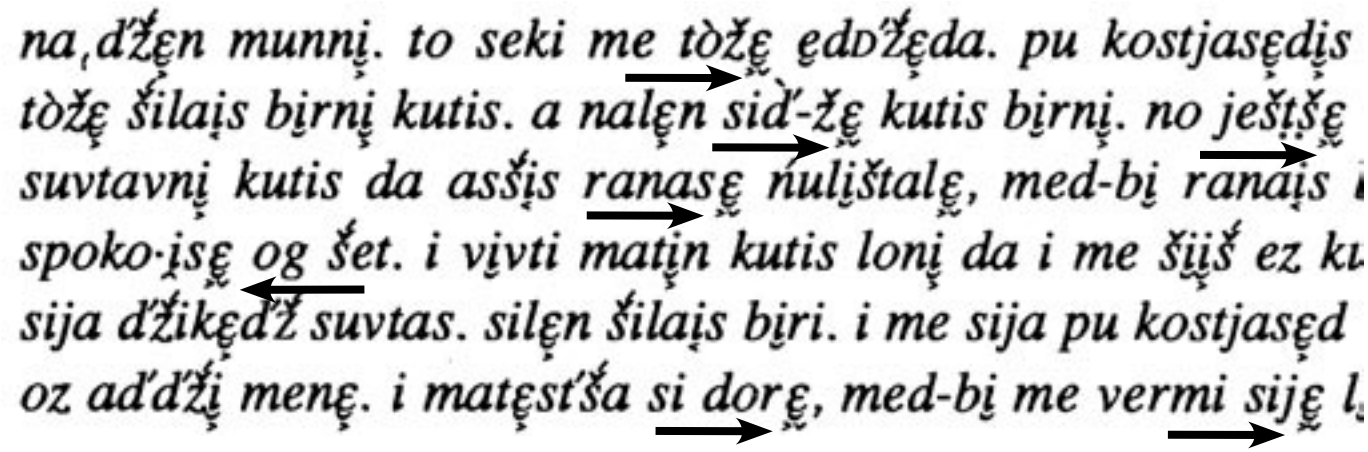


Figure 5: Syrjänische Texte IV. Gessammelt von T. E. Uotila. Übersetzt und herausgegeben von Paula Kokkonen. — MSFOu 221. Helsinki 1995.

This combining diacritical mark should be added with the same rationale as the existing 0356 COMBINING RIGHT ARROWHEAD AND UP ARROWHEAD BELOW.

3. Two pre-UPA characters

In the 19th century the Swedish landsmålsalfabetet developed for the transcription of Swedish dialects and the corresponding Norwegian transcription scheme, Norvegia, were in use in Swedish and Norwegian Uralistics as well (the UPA transcription scheme was established in the beginning of the 20. century; e.g. Sovijärvi 1965). Figure 6 shows a definite statement on this pre-UPA use of the Swedish landsmålsalfabetet in a Lule Sámi Dictionary. In his classical research on the Nordic loanwords in the Sámi languages Qvigstad (1893) makes a corresponding statement according the use of Norvegia.

Being pre-UPA transcription schemes the landsmålsalfabetet and Norvegia matter also from the Uralists' point of view, which means that the inclusion of the characters used by the Scandinavian transcription schemes (landsmålsalfabet, Norvegia and Dania) benefits the Uralists' community as well. Hopefully a proposal for those alphabets can be produced in the near future.

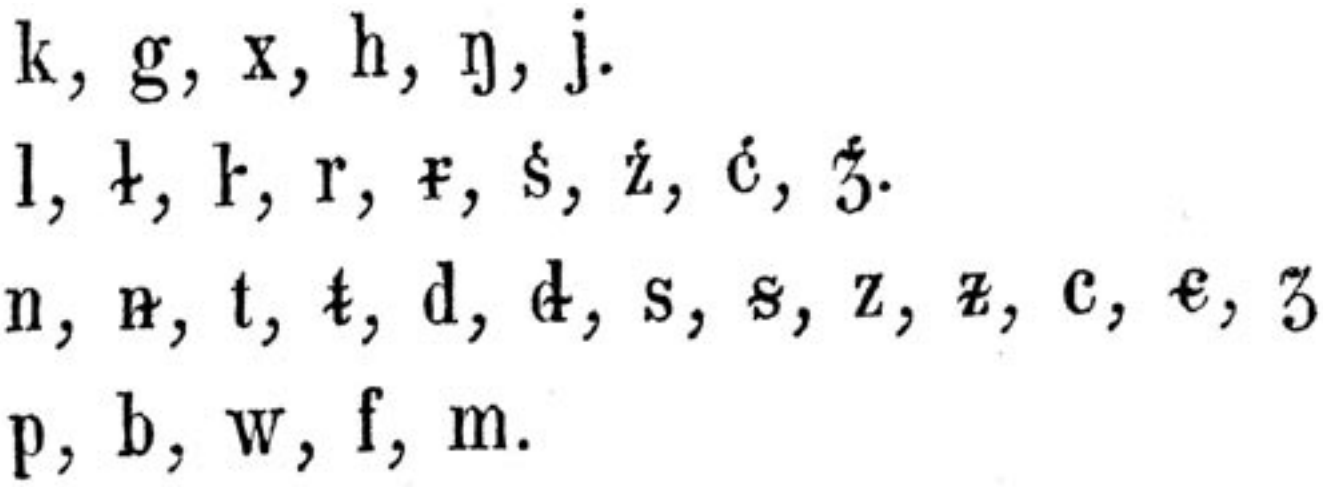
Von den hier bei der schreibung der lappischen wörter gebrauchten typen des schwedischen dialektalphabetes (»det svenska landsmålsalfabetet») sind

a = schwed. **a** in **kasta**, finn. **a**.
c = **ts**; *c̣* = **tts**.
č = schwed. **tj**; *č̣* = **ttj**.
ʒ = tönendes **dd** + tonloses **s**.
ǯ = tönendes **dd** + das tonlose **j** in schwed. **tjära** (also nicht tönendes **j** wie in engl. **judge**).
đ = mouilliertes **d**.
e = **e** in schwed. **ren, hem**, deutsch **lehm**.
ə = ein unvollkommener, ö-ähnlicher vokal.
æ = ein wenig offener als das ä in schwed. **tjäna, tjänst**.
a = ä in schwed. **ära, ärt**, finn. **härkä**.
g = (hinteres, »gutturales») **g** in schwed. **gård**, d. **gabe**.
g = (vorderes, »palatales») **g** in schwed. **snigel**, d. **geben**.
i = **i**; *j* = **j**.
k = (hinteres, »gutturales») **k** in schwed. **karl**, d. **kauen**.
ķ = (vorderes, »palatales») **k** in schwed. **icke**, d. **kehren**.
ŋ = mouilliertes **n**.
g = (hinteres, »gutturales») **ng** in d. **lang**.
ŋ = (vorderes, »palatales») **ng** in d. **eng**.
o = **o** in schwed. **komma**, finn. **otan**.
r = gerolltes zungenspitzen-r.
š ungef. = schwed. **sj**, d. **sch**.
u = finn., d. **u**.
v = **v**.
w = engl. **w**.
h = tonloses engl. **w**.
ɔ = tonloser vokal (das ende des vorhergehenden vokales ist tonlos).

Ein $^{a\ e\ ə\ o\ u}$ bezeichnet einen sehr kurzen svarabhakti-vokal von ungefähr derselben qualität als resp. *a*, *e*, *ə*, *o*, *u*. Nach triftongen ist dieser svarabhakti-vokal oft nicht hörbar; man hört z. b. sowohl *ru̯əu^{e}tę* als *ru̯əutę*.

Figure 6: K. B. Wiklund: Lule-lappisches Wörterbuch. — MSFOu 1. Helsinki 1890.

However, Finnish scholars of that time used a different pre-UPA transcription scheme as the example in Figure 7 shows:


k, g, x, h, ŋ, j.

l, ɬ, ɭ, r, ɾ, ś, ź, ć, ʒ́.

n, ɲ, t, ŧ, d, ɖ, s, ʂ, z, ʐ, c, ȼ, ʒ

p, b, w, f, m.

Figure 7: M. A. Castrén: Grammatik der samojedischen Sprachen. St. Petersburg 1854.

In connection with the work for a proposal for the characters of the Swedish, Norwegian and Danish schemes this Finnish pre-UPA transcription scheme should also be taken into account.

Independently from such a proposal the Álgu project indicates a need for the addition of two pre-UPA characters — which according to our knowledge — are not part of the schemes mentioned above but nevertheless seem to be invented following the principles of the Swedish landsmålsalfabetet and Norvegia. The characters in question are used in the main printed source of the Álgu project for the Pite Sámi language (Halász 1896), a language belonging to the smallest and least documented Sámi languages (Figure 8).

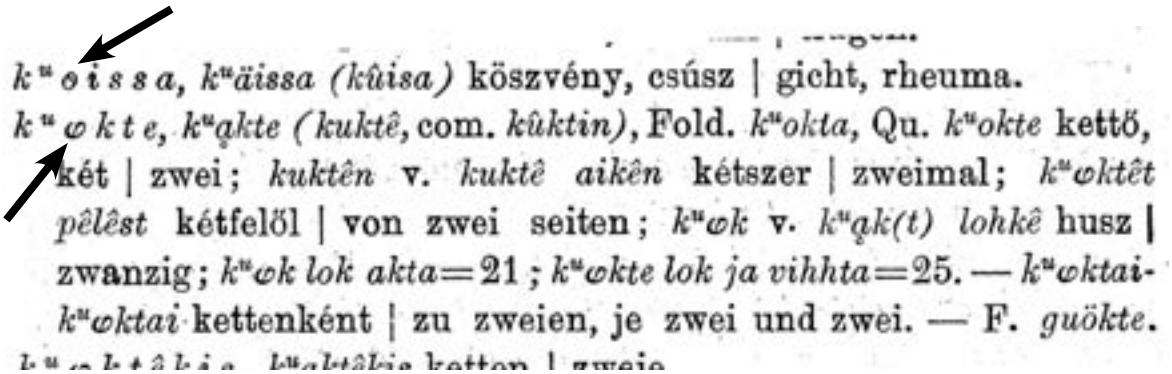

kᵘ o i s s a, kᵘäissa (kûisa) köszvény, csúsz | gicht, rheuma.
kᵘ ɷ k t e, kᵘąkte (kuktê, com. *kûktin),* Fold. *kᵘokta,* Qu. *kᵘokte* kettő, két | zwei; *kuktên* v. *kuktê aikên* kétszer | zweimal; *kᵘɷktêt pêlêst* kétfelől | von zwei seiten; *kᵘɷk* v. *kᵘąk(t) lohkê* husz | zwanzig; *kᵘɷk lok akta*=21; *kᵘɷkte lok ja vihhta*=25. — *kᵘɷktai-kᵘɷktai* kettenként | zu zweien, je zwei und zwei. — F. *guökte.*

Figure 8: Halász Ignácz: Pite Lappmarki szótár és nyelvtan. Budapest 1896.

As mentioned above RILF plans to make the Álgu database publicly available as a scientific archive in a not so distant future. However, the condition for this is consistency of data and character encoding. For this reason we need to bring up those two pre-UPA characters used by Halász in the same proposal together with the UPA characters mentioned above in part 2, although the project is now able to progress the work with some interim solutions (by using forged encodings), not suitable for publication, though.

4. Lexicographic symbols

The following symbols we want to bring up in this paper in order to ask the experts of the working group for advice.

4.1 Roman numeral superscripts
How roman numerals as superscripts should be dealt with? The question arose in connection with the digitalization of the dictionary of the Karelian language. In this dictionary homonyms are marked by superscripted Roman numerals as in Figure 9:

sorevuo v. → sorie[II]. **1.** inkoat. *vihmah sorevui. vezi šorevui vuodamah.* Suoj | *sorevui vihmah. vihmaš šorevui, pilvi nouzi.* Säämäj **2.** pass. *sorevuo ~ šorevuo.* Suoj | *sorevuo ~ šorevuo. sorevui kazvoi tobjakse tšomaste. kazvamaš šorevui vihmuhuv vilad.* Säämäj

sorie[I] v. selvittää, oikoa. *šorie.* Uhtua Tulemaj | *langoi soritah, poaznoi eroitellah viibes.* Suoj | *langoida šoritah viipsihyöh, štobi hiän šeľľitťäätšiis (langad yhteh tartunnuot).* Tver | *šoria.* (panna kangas puille) Impil, Vitele | *sorie.* Salmi | *nuonat soritah.* Säämäj ‖ Kuv. *ažeida soritah hyväh loaduh.* Suoj | *dieluo soritah pa(j)istah ilmai kiistämättäh.* Säämäj

sorie[II] v. **1.** sataa kaatamalla; lorista, lotista; porista. *đo vezi šoriu, rubieu kiehumah.* Jyskyj | *vihmuo sorie*

Figure 9: Karjalan kielen sanakirja. 1–6. — Lexica Societatis Fenno-Ugricae XVI. Helsinki 1968–2005.

For the time being there is 1D35 MODIFIER LETTER CAPITAL LETTER I, which can be used for numbers up to 3, but it is a letter not a numeral. The correct correspodences for the numeral superscripts would be the Roman numerals in the code places 2160…2180. Is this a matter of encoding or not?

4.2 Symbols for non-existence or conditional existence
The comprehensive dictionaries of the dialects of Lule and Northern Sámi use a series of symbols indicating that a certain word is not known in an area or by an informant. The symbols differ according to the grade of non-existence as shown in Figure 10:

indicates that the meaning of a word (in every case or in a particular sense) is unknown.

# indicates that the meaning of a word is known, but that it does not belong to the dialect in question, or that the native assistant had not heard the word used in that particular sense, though he would not go so far at to say that it could not be understood.

† indicates that the word is formed in accordance with the rules of the dialect, but is not considered by the native assistant to be a current word. In the corrections at the end of the volume this sign means that something is to be deleted.

Figure 10: Konrad Nielsen: Lapp Dictionary. 1–5. — Instituttet for sammenlignende kulturforskning, Serie B: Skrifter XVII. Oslo 1932–1962. [2. ed. 1979.]

Figure 11 shows an example: The meaning of the verb boalbâdit is not known in Kautokeino (Kt) and in Polmak (P) it is not recognized by all individuals. The adjective boal'bai is not recognized to be a current word in Karasjok and Kautokeino.

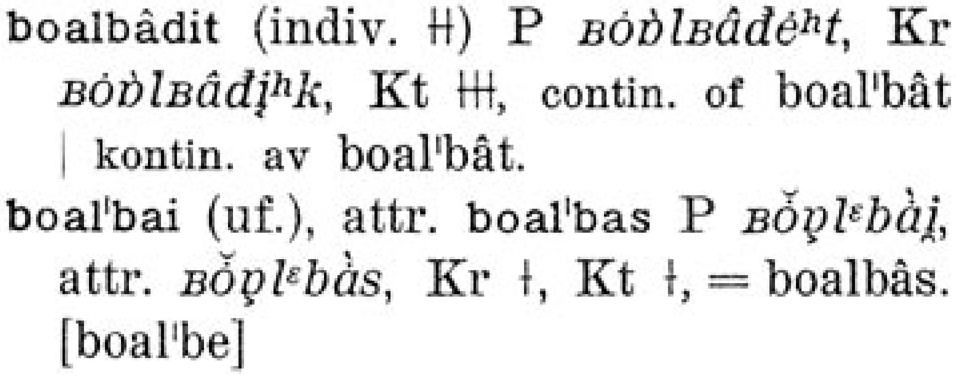
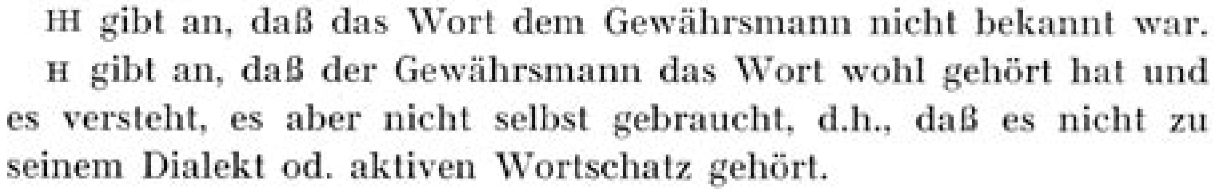

boalbâdit (indiv. Ħ) P ʙóὸlʙâdėʰt, Kr ʙóὸlʙâdįʰk, Kt ĦĦ, contin. of boal'bât | kontin. av boal'bât.
boal'bai (uf.), attr. boal'bas P ʙŏϱlᵋbàį, attr. ʙŏϱlᵋbàs, Kr ꟷ, Kt ꟷ, = boalbâs. [boal'be]

Figure 11: Konrad Nielsen: Lapp Dictionary. 1–5. — Instituttet for sammenlignende kulturforskning, Serie B: Skrifter XVII. Oslo 1932–1962. [2. ed. 1979.]

For areal and historical linguistics those symbols are very useful. Examples from the Lule Sámi dictionary show a corresponding use of the symbols in question (Figure 12 and 13):


ĦĦ gibt an, daß das Wort dem Gewährsmann nicht bekannt war.
Ħ gibt an, daß der Gewährsmann das Wort wohl gehört hat und es versteht, es aber nicht selbst gebraucht, d.h., daß es nicht zu seinem Dialekt od. aktiven Wortschatz gehört.

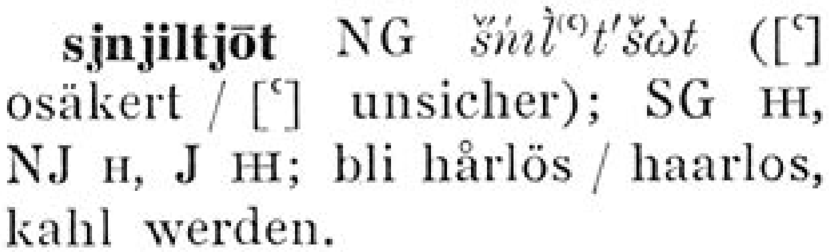

sjnjiltjōt NG š̌ńì̀l'ᶜt'š̌ὼt ([ᶜ] osäkert / [ᶜ] unsicher); SG ĦĦ, NJ Ħ, J ĦĦ; bli hårlös / haarlos, kahl werden.

Figure 12 & 13: Harald Grundström: Lulelappisches Wörterbuch. 1–4. — Schriften des Instituts für Mundarten und Volkskunde in Uppsala, Reihe C: 1. Uppsala 1946–1954.

As far as we understand these symbols are not yet encoded in the UCS. Possibly the use of these symbols is wider than indicated here. Feedback from the experts of the working group would be very welcome.

Literature

Antti Iivonen & Antti Sovijärvi & Reijo Aulanko 1990: Foneettisen kirjoituksen kehitys ja nykytila. — Mimeographed Series of the Department of Phonetics, University of Helsinki 16.

Antti Sovijärvi & Reino Peltola 1965: Suomalais-ugrilainen tarkekirjoitus. — Publicationes Instituti Phonetici Universitatis Helsingiensis 9.

J. K. Qvigstad 1893: Nordische Lehnwörter im Lappischen. — Christiania Videnskabs-Selskabs Forhandlingar for 1893 No. 1. Christiania.

MSFOu = Mémoires de la Société Finno-Ougrienne.

Appendix

As can be observed the concept of UPA differs from the one underlying the Scandinavian schemes and also IPA. In the latter letters were modified whereas in UPA the basic letters remain in general untouched and for variations combining diacritical marks are used. This means the UPA transcription scheme is an open one. With basic letters can be combined multiple diacritics as long as the result is phonetically sensible. Different linguists have used it in different ways (e.g. excessively or simplified) depending from the purpose. Due to the openess of the scheme it is not surprising that some characters were looked over in the working process for the original UPA proposal in 1999.

As an illustration for the openess of the system it can be mentioned that to a very large extend basic letters with diacriticals attached are used as Modifier Letters. Some of them are Modifier Letter variants of precomposed letters as follows:

1D43+0308 ᵃ̈ — 00E4 ä
1D52+0308 ᵒ̈ — 00F6 ö
1D43+030A ᵃ̊ — 00E5 å
1D2C+030C ᴬ̌ — 00C4 Ä
1D49+0307 ᵉ̇ — 0117 ė
1D52+0307 ᵒ̇ — 022F ȯ
1D58+0308 ᵘ̈ — 00FC ü
02E2+0301 ˢ́ — 015B ś
02E2+030C ˢ̌ — 0161 š

However, we do tend to think that — at least at this stage — we should not try to add those into the UCS, since they can sufficiently be presented through combination — if only the font supports these combinations.

Even Modifier Letter variants of not precomposed characters were found:

02E2 MODIFIER LETTER SMALL S with the following diacritical marks:
030C+0301 ◌̌◌́
1D2C+0355 ᴬ◌͕
1D4C+0311 ᵀ◌̑

The new survey made in 2005 contains the statistics of almost 1,500 lines of text written in UPA transcription including material from the following languages: URALIC: Finnish, Ingrian, Karelian, Ludian, Vepsian, Votic, Estonian, Livonian, Sámi languages (10), Erzya, Moksha, Mari, Komi, Udmurt, Mansi, Hanti, Hungarian, Nenets, Nganasan, Selkup, Kamass; TURKISH: Chuvash, Misher, Kumyk; MONGOLIAN: Mongol, Oirat; TUNGUS: Eveni; Ket.

As mentioned above a result of this survey is a statistics of frequency of the combination of basic letters and combining diacritical marks. This material will be made publicly available e.g. for the use of font designers.