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 add LATIN CAPITAL LETTER N WITH LONG RIGHT LEG to the UCS.**
Source: **Michael Everson**
Status: **Expert Contribution**
Date: **2000-11-28**

A. Administrative

1. Title
Proposal to add LATIN CAPITAL LETTER N WITH LONG RIGHT LEG to the UCS.
2. Requester's name
Michael Everson.
3. Requester type
Expert contribution.
4. Submission date
2000-11.29
5. Requester's reference
None.
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. Latin Extended-N
2. Number of characters
1.
3. Proposed category
Category A.
4. Proposed level of implementation and rationale
Level 1. Base character with no diacritics.
5a. Character names included in proposal?
Yes.
5b. Character names in accordance with guidelines?
Yes.
5c. Character shapes reviewable?
Yes:

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 below).
7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?
No.
8. Does the proposal address other aspects of character data processing?
No.

C. Technical -- Justification

1. Contact with the user community?
No, but their publications are well-known.
2. Information on the user community?
See below.
3a. The context of use for the proposed characters?
Completes a case-pair with U+019E LATIN SMALL LETTER N WITH LONG RIGHT LEG; used in popular Lakota orthography to indicate the nasality of a preceding vowels.
3b. Reference
See below.
4a. Proposed characters in current use?
Yes.
4b. Where?
In North America.
5a. Characters should be encoded entirely in BMP?
Yes.
5b. Rationale
Keeping them with other Latin characters used by Lakota.
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. It is similar, but not identical, to LATIN LETTER ENG
7b. Where?
7c. Reference
8a. Can any of the characters be considered to be similar (in appearance or function) to an existing character?
No.
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.

D. Proposal

On the Unicode list, Ken Whistler recently provided further discussion of the issue of Unicode coverage of the Lakota orthography. I reproduce some of his points here.

> The issue for Lakota in Unicode is the representation of the Lakota nasal vowels in the 1982 Lakota orthography. That orthography was developed by Lakota educators, was adopted by the South Dakota Association of Bilingual and Bicultural Education, and is being used to print books, dictionaries, and teaching materials for Lakota.
>
> Lakota has three nasal vowels, a nasalized form of /i/, /a/, and of /u/. The 1982 orthography indicates these with digraphs, where the second element is basically an n with a long right leg. Earlier discussion of this had pointed to Unicode U+019E LATIN SMALL LETTER N WITH LONG RIGHT LEG as this character. But that character has no associated uppercase character, which is needed for the Lakota orthography.
>
> The issue is complex, however. It is clear that this Lakota letter is a new creation. If you go back to the source of this element of the orthography, you can find it in Buechel, 1939, *A Grammar of Lakota*, which represents the vowels this way, but using what is clearly a lowercase Greek letter ETA (i.e. U+03B7). This, in turn, derived from a 19th century Dakota alphabet created by Episcopal missionaries and associated particularly with the name of Stephen R. Riggs. The Greek letter ETA was often a printing substitution for ENG (i.e. U+014B), to indicate nasalization. So we have a complicated confusion here of three letterforms.

This, of course is nothing new to the UCS. We are familiar with such pairs:

Ð	U+00D0	LATIN CAPITAL LETTER ETH
ð	U+00F0	LATIN SMALL LETTER ETH
Đ	U+0110	LATIN CAPITAL LETTER D WITH STROKE
đ	U+0111	LATIN SMALL LETTER D WITH STROKE
Ɖ	U+0189	LATIN CAPITAL LETTER AFRICAN D
ɖ	U+0256	LATIN SMALL LETTER D WITH TAIL

Other traditions have also confused GREEK SMALL LETTER ETA with LATIN SMALL LETTER ENG. Chief among them is the Uralicist community, whose Finno-Ugric Phonetic Alphabet (FUPA) made use of GREEK SMALL LETTER ETA which was available in early-twentieth-century fonts. But this is complicated by the modern use of LATIN SMALL LETTER ENG in ordinary Sami orthography, as well as in IPA and FUPA documentation.

> U+019E was proposed in the IPA *Principles* (1949) for use in digraphic spellings of nasal vowels -- presumably as a way of regularizing the ETA/ENG confusion. But the letter was withdrawn from the IPA in 1976.

Actually, the 1949 *Principles* state two uses of the character:

> *Clause 28:*
> Japanese syllabic nasal: **ŋ**.
>
> *Clause 29.k:*
> To represent nasalised vowels; for instance it may sometimes be found convenient (especially in phonetic orthography) to write a**ŋ** or **aŋ**, **ɛŋ** or **ɛŋ**, in place of …**ã**, **ɛ̃**, etc.

> However, presumably because of the enormous impact of the missionary orthography on the history of the written Lakota language, the digraphic spelling of nasal vowels was preferred by the Lakota educators when deciding on the 1982 orthography, over the general Siouan linguistic tradition of writing nasal vowels with ogoneks. Effectively, this meant a resurrection of the N WITH LONG RIGHT LEG, since the orthography was intended to be Latin, not Latin with one Greek letter ETA.

This is *precisely* the reason I have proposed, for several years, the disunification of *CYRILLIC LETTER KU* from LATIN LETTER Q and *CYRILLIC LETTER WE* from LATIN LETTER W. Kurdish orthography is intended to be Cyrillic, not Cyrillic with two Latin letters.

> The practical orthographies used in the missionary dictionaries and grammars, and technical linguistic orthography of Boas and Deloria never had to decide on the problem of how to uppercase the nasal vowel, since as a digraphic representation, the nasal indicator never occurs initially, and those sources don't use all-cap text anywhere. But the 1982 orthography is intended for general use-- and that means that the Lakota text can also occur in all-cap environments such as chapter headers, and so on.

I expect that this means that the Lakota name of the film *Dances with Wolves* would be written something like the following (Siouan linguistic orthography given last).

Ṡuŋk manitu ṫaŋka ob waci
ṠUŊK MANITU ṪAŊKA OB WACI
ṠUŊK MANITU ṪAŊKA OB WACI
Šųk manitu t‘ąka ’ob wači

(More or less literally, this parses as ‘dog-of-wilderness sacred together-with he-dances’.)

> So as in the case of African languages that adopted an IPA-based orthography, and then created uppercase versions of letters that had no uppercase in IPA (*cf.* U+0186, U+018F, U+01A9, for example), we have another instance here of orthographic usage driving the need for a new uppercase character: LATIN CAPITAL LETTER N WITH LONG RIGHT LEG.

All things considered, I support this suggestion. But I want to bring something up for discussion. It is known that the LATIN CAPITAL LETTER ENG has been found with more than one glyph variant. Apparently the round one may still have currency in Africa, though the angular one, which is preferred in Sami orthography, is used as the paradigmatic letter shape in ISO/IEC 10646 and the Unicode Standard. The question which we should ask is whether the new Lakota letter admits of similar glyph variation.

Ŋ/ŋ	U+014A	LATIN CAPITAL LETTER ENG
ŋ	U+014B	LATIN SMALL LETTER ENG
Ɲ	U+019D	LATIN CAPITAL LETTER N WITH LEFT HOOK
ɲ	U+0272	LATIN SMALL LETTER N WITH LEFT HOOK
ɳ/Ɲ	*proposed*	*LATIN CAPITAL LETTER N WITH LONG RIGHT LEG*
ɳ	U+019E	LATIN SMALL LETTER N WITH LONG RIGHT LEG

Ɲ	*?*	*LATIN CAPITAL LETTER N WITH RETROFLEX HOOK*
ɳ	U+0273	LATIN SMALL LETTER N WITH RETROFLEX HOOK
ɲ	*proposed*	*LATIN SMALL LETTER N WITH CURL*
Η	U+0397	GREEK CAPITAL LETTER ETA
η	U+03B7	GREEK SMALL LETTER ETA

I'm almost tempted to propose capital forms for all such Latin letters, such as *LATIN CAPITAL LETTER N WITH RETROFLEX HOOK* – merely to forestall the creeping in of new letters from time to time, but I am certain that eminent standardizers like Ken Whistler would gnash their teeth crying "Ɲɲɲ!" were I to do so.

But the question is, I suppose, for the Lakota; would the forms given below ever occur?

ṠUŊK MANITU ṪAŊKA OB WACI

ṠUŊK MANITU ṪAŊKA OB WACI

It is possible to find parallel forms in Sami, though the forms given on the left are preferred. These are placenames.

Áŋŋel	Jiekŋaáhpi
ÁŊŊEL/ÁŋŋEL	JIEKŊAÁHPI/JIEKŋAÁHPI
AŊŊEL/AŋŋEL	JIEKŊAÁHPI/JIEKŋAÁHPI