N3812

TO: Mike Ksar, ISO/IEC JTC 1/SC2 WG2 Convener

FROM: Hans-Jörg Bibiko, Max Planck Institute for evolutionary Anthropology, Department of Linguistics

RE: On the proposed U+A78F LATIN LETTER MIDDLE DOT (L2/09-031R = N3567)

DATE: 7 April 2010

L2/10-118

Dear Mr. Ksar,

1. Background. I've been working since years at the Max Planck Institute for evolutionary Anthropology, Department of Linguistics, as computer scientist mainly focused on the general text processing – esp. texts encoded in ISO 10646 – for linguistic research of any kind. In my daily work I'm very often confronted with encoding issues of characters used by thousands of languages as in for instance "The World Atlas of Language Structures" by Haspelmath et al., 2005, where we gathered data from 2,560 languages.

One of the main issues is the chance to write the same character (graphical representation) by using different ISO 10646 code points as in IIII U+0049 U+0399 U+0406 U+4C0. Algorithms which wants to analyze texts has to know that each of these ISO 10646 code points could stand for a 'I' according to the underlying script or language. It is quite often the case that authors simply forgot to switch their computer input method to e.g. Greek to type a 'I'.

Due to this background I reviewed the proposal about to encode a Middle Dot letter for Phags-pa transliteration (N3567, N3568, N3694, and N3694) by Andrew West.

2. General remarks. Mark Davis says in (http://www.unicode.org/standard/WhatIsUnicode.html):

"Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language."

Due to multifarious reasons – which I'm aware of – the standard ISO 10646 has not remained quite as simple as it originally planned to provide a unique number for every character.

The main question is whether ISO 10646 should distinguish between characters and functions of a character in a given language or script respectively. For instance the character '!' has amongst others two main code points EXCLAMATION MARK (marked as punctuation) and LATIN LETTER RETROFLEX CLICK (marked as general sign used by languages). Here the function was clearly differentiated. This distinction leads to a problem that each user who is using the ISO 10646 encoding has to be aware of it. In the case for '!' many authors make usage of the character EXCLAMATION MARK if they are writing for instance Khoisan languages simply because it is easier to type on a computer keyboard. In other words the more options an user has to type a character the more problems arise especially for text analyzes done by computer software afterwards.

3. Encoding of Middle Dot.

U+00B7 MIDDLE DOT

This character could be used as alternative for the proposed U+A78F LATIN LETTER MIDDLE DOT but it is marked as punctuation sign. Programmers of software applications often rely on the ISO 10646 specifications but unfortunately not always. (See e.g. N3678 '3. Search Issues for U+00B7 MIDDLE DOT').

The only chance to use MIDDLE DOT is to change the ISO 10646 property of that code point to not being marked as punctuation sign. Since this code point is well-established for years this is rather unlikely.

U+02D1 MODIFIER LETTER HALF TRIANGULAR COLON

This character fulfills all Andrew West's criteria as alternative for the proposed U+A78F LATIN LETTER MIDDLE DOT, except for its shape. Originally this character was displayed as filled circle but its shape changed over the time. This issue can be solved – if it would turn out that this could lead to any problems – by

using different fonts easily (as for the Latin letter ‘g’ there are at least two different ways to visualize it: ‘g’ and ‘ɡ’).

4. Summary. The insertion of the proposed U+A78F LATIN LETTER MIDDLE DOT as the middle dot for Phags-Pa is needless and would increase the likelihood of choosing the wrong character by users and mismatches of computational text analyzes. Furthermore the insertion of LATIN LETTER MIDDLE DOT could govern to add new ISO 10646 code points of other look-alike characters.

The usage of U+02D1 MODIFIER LETTER HALF TRIANGULAR COLON as middle dot is normally accepted amongst a wide range of linguists.