ISO/IEC JTC1/SC2/WG2 N2976
L2/05-172R
2005-08-26

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 clarify principles for disunification of combining diacritical marks**
Source: **USA (INCTS/L2) and the Unicode Technical Committee**
Status: **Member Body and Liaison Contribution**
Date: **2005-08-26**

Summary
This document presents a proposed set of criteria for deciding when to encode separate script-specific diacritical marks. The US Member Body and the UTC are requesting that these criteria be added to the WG2 document on Principles and Procedures. In making the case for the proposed criteria, this document provides additional background information, as well as a discussion of the purpose of the "common" diacritical marks and a brief mention of security related issues.

Background
The Unicode Standard states about the Combining Diacritical Marks block: "The combining diacritical marks in this block are intended for general use with any script." This text is sometimes misunderstood as if it was intended to be a normative directive that these diacritical marks should be used with all scripts.

It is indeed true that in the ISO/IEC 10646 and in the Unicode Standard every diacritical mark may be applied to any base character—but this does not imply, or require, that doing so will lead to a graphically meaningful result, or that any particular combination of base character and diacritic will be supported by applications. It is merely a general principle about the use of the standard, namely that such sequences are not illegal as they would be in ISO/IEC 6937.

The Universal Character Set has encoded many script-specific combining marks, even where they bear a superficial graphical similarity to a generic diacritical mark. Recently encoded examples include U+0659 ◌ٙ ARABIC ZWARAKAY, U+065A ◌ٚ ARABIC VOWEL SIGN SMALL V ABOVE, U+065B ◌ٛ ARABIC VOWEL SIGN INVERTED SMALL V ABOVE, and U+065C ◌ٜ ARABIC VOWEL SIGN DOT BELOW, which were not unified with U+0304 ◌̄ COMBINING MACRON, U+030C ◌̌ COMBINING CARON, U+0302 ◌̂ COMBINING CIRCUMFLEX ACCENT, and U+0323 ◌̣ COMBINING DOT BELOW, respectively.

Likewise, U+135F ◌፟ ETHIOPIC COMBINING GEMINATION MARK was not unified with U+0304 ◌̈ COMBINING DIAERESIS, and U+05C5 ◌ׅ HEBREW MARK LOWER DOT was not unified with U+0323 ◌̣ COMBINING DOT BELOW.

And the long-encoded U+05B5 ◌ֵ HEBREW POINT TSERE, U+0738 ◌ܸ SYRIAC DOTTED ZLAMA HORIZONTAL, as well as the more recently encoded U+0CBC ◌಼ KANNADA SIGN NUKTA, have all been distinguished from U+0324 ◌̤ COMBINING DIAERESIS BELOW despite the fact that in appearance they

simply consist of two side-by-side dots placed below a character. A superficial graphic similarity was not considered sufficient reason to justify unification.

Purpose of the "Common" Combining Diacritical Marks

The combining diacritical marks in the two blocks in the standard for generic combining marks are primarily for use with a number of European scripts, namely, Greek, Latin, Cyrillic, and Georgian. These scripts are alphabets, and share a general typographical model, including the common application of diacritic marks to indicate accents and pronunciation modifications of letters. Adaptations of these scripts for specialized notational systems—for instance, phonetic alphabets or Western mathematics—and for orthographies of non-European languages, also make heavy use of these combining diacritical marks.

While other scripts, notably the Arabic script, also make very heavy use of diacritical marks and other kinds of annotation marks, scripts not directly derived from Greek have their own history of diacritic development. Therefore the various dots and other marks used with them cannot automatically identified with "common" combining diacritical marks based on graphic form alone. The exception to this rule would be a case where a common diacritical mark has been explicitly borrowed (usually from the Latin script) for use in an unrelated script.

In the general case it would be wrong to presume that the application of common diacritical marks will make much sense (or be reasonably supported by applications) when applied to characters from different typographical traditions, such as CJK ideographs or Sumero-Akkadian syllables, for example. The determination of what are "reasonable" combinations should be guided, in large part, by established typographical practice for each script. The proposed criteria are intended to help guide this decision making process.

Security Issues

There are security issues involved in any disunification. Unicode Technical Report #36, Unicode Security Considerations, discusses many details of such issues.

The US Member Body and the UTC request that the following text be added to the Principles and Procedures document:

> **Criteria for disunification of combining diacritical marks**
>
> A number of criteria may be considered when deciding whether a proposed combining diacritical mark for a particular script should be unified with an existing encoded combining diacritical mark.
>
> The main criterion is that of the shape of the glyph, since that is the chief identifier of a diacritical mark. When the range of glyphic appearance of a diacritical mark may be markedly different from the range typical of the generic diacritical mark, disunification may be preferred. These and other criteria have been used in the past, and may be used in the future, as deciding factors in whether to encode separate diacritical marks (or to disunify) for particular scripts. Among other criteria which would favour a decision to disunify when encoding are:

a. the mark forms part of a set of marks in the script (for example a set of tone marks), but only some members of the set could be considered candidates for unification with existing marks.

b. the mark has a specific function fundamentally unrelated to the generic diacritical mark—for instance, the use of the mark as a vowel sign as opposed to the use of a similar-shaped mark as a modifying diacritic. In such case the two uses might also require explicit differences in their character properties.

c. the display behaviour is fundamentally different and requires different support. For example, U+A806 SYLOTI NAGRI SIGN HASANTA looks like a combining circumflex, but requires different display support.

d. the mark has been borrowed from another script, but has been significantly modified to fit with the ductus of the borrowing script, disunification may be preferred.

The more of these criteria are satisfied, and the stronger the degree to which each is satisfied, the stronger the case for encoding a script-specific diacritical mark. This is not a matter of a rule that deterministically yields a "yes/no" decision; rather, it is a question of degree, which can then form a basis for a proper judgement of the encoding question.

In general, these criteria are not much different from those used for assigning script-specific punctuation.