

Title:	Response to the revised "Final proposal for encoding the Phoenician script in the UCS" (L2/04-141R2)
Source:	Peter Kirk
Status:	Individual Contribution
Action:	For consideration by the UTC
Date:	2004-06-07

This paper is a response addressed to the UTC to the revised "Final proposal for encoding the Phoenician script in the UCS", ISO/IEC JTC1/SC2/WG2 N2746R2 and L2/04-141R2, <http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2746>. It also includes recommendations concerning the "review of the factors that we should take into account in determining whether to unify two scripts or not, to make sure that they make sense for historic scripts [which] is already on the agenda for the next UTC meeting in June" (quoting Mark Davis writing on 2004-05-11, <http://www.unicode.org/mail-arch/unicode-ml/y2004-m05/0719.html>).

Note: This response is in addition to this respondent's comments to the UTC made on 2004-04-29 on the original version of the proposal; it is noted that his point that the answer to question C2a was misleading has been addressed in the revised proposal.

Summary

The proposal to encode Phoenician as a separate Unicode script has been highly controversial. The proposal is based on the proposer's understanding of the Phoenician set of letters (an *abjad*) as a distinct set of abstract characters. However, it seems to be the general view of scholars of Semitic scripts that there is a single north-west Semitic script with 22 characters, and that Phoenician letters, square Aramaic/Hebrew letters, and several other letter styles are variant glyphs for these same characters. For this reason a number of scholars of Semitic languages have rejected the proposal on the grounds that it violates the Unicode design principle of encoding characters, not glyphs. Other Semitic scholars have accepted the proposal, but it seems likely that they do not understand the Unicode character-glyph model, rather than that they dissent from the general scholarly opinion that there is a single set of north-west Semitic characters.

Nevertheless, there is a significant demand for separate encoding of Phoenician letter forms. However, it has not been clearly demonstrated that there is a real need, rather than just a desire, for these letters to be distinguished *in plain text* from the existing Unicode Hebrew characters.

This response can be summarised as a recommendation that the UTC should first complete its proposed review of criteria for unifying historic scripts and reaffirm the principle that sets of glyph variants of existing characters should not be encoded as separate scripts; and then, when considering the Phoenician script proposal, should look at the following questions in order:

1. Are the proposed Phoenician characters in fact distinct abstract characters, or are they glyph variants of characters which are already encoded in the Unicode Hebrew block?
2. If Phoenician letters are not distinct abstract characters, is there an exceptional need to distinguish them *in plain text* from the existing Hebrew characters?
3. If there is such an exceptional need, what is the appropriate mechanism for making the distinction?

Additionally, the proposed name of the new characters is misleading. To avoid further confusion between language, script and glyphs, the Unicode name for these glyphs should be not "Phoenician" but a more generic term. The suggested name "Old Canaanite" is a good one.

Background

The scholarly understanding of north-west Semitic script

From late April until early June 2004 there has been wide-ranging and sometimes acrimonious discussion about the original version of this proposal, on the Unicode and ANE mailing lists and elsewhere. This discussion has clearly demonstrated that the proposal is unacceptable to a significant section of the community of users of the Phoenician form of writing. The major issue has been over the principle of whether Phoenician should be encoded as a separate script from Hebrew. This has been asserted by the proposer, but it has been vehemently disputed by some scholars of Semitic languages. According to the Unicode Standard 4.0.0 section 1.1 (p.3), "the Unicode Standard must also respond to scholarly needs". But there is little indication that scholarly needs and understandings, as expressed by scholars of Phoenician and related writing and languages who have responded to the proposal, have been taken into account in this proposal.

The reactions on various lists suggest that most of those who have a scholarly interest in texts written with these glyphs consider Phoenician letters to be glyph variants of Hebrew letters, not distinct characters. To quote Patrick Durusau, Director of Research and Development of the Society of Biblical Literature, writing to the Unicode list on 2004-05-24 (<http://www.unicode.org/mail-arch/unicode-ml/y2004-m05/1375.html>), "long PRIOR to Unicode, Semitic scholars reached the conclusion all Semitic languages share the same 22 characters." (Durusau in fact supports the proposal, apparently because he sees a distinction between the scholars' and the Unicode concepts of "character"; "all Semitic languages" should in fact read "all north-west Semitic languages".) Others have used stronger language: one internationally renowned scholar, Dr Stephen A. Kaufman, wrote on 2004-05-01: "Anyone who thinks there has to be a separate encoding for Phoenician either does not understand Unicode or (and probably "and") does not understand what a glyph is" (quoted from <https://listhost.uchicago.edu/pipermail/ane/2004-May/012945.html>). (Kaufman misunderstands the situation, because to him the proposer's position that Phoenician glyphs are separate abstract characters is inconceivable.) Scholarly usage reflects this understanding. Inscriptions originally written with Phoenician style glyphs are routinely represented in print with square Aramaic/Hebrew glyphs (i.e. like those used as reference glyphs for the Unicode Hebrew block). These are not considered to be transliterations, but to be faithful representations of the original text with more widely recognisable glyphs.

To put this in Unicode terms, this evidence clearly indicates that, in the view of the Semitic scholars who are the experts in the field, the north-west Semitic languages are written all with the same set of 22 abstract characters. Therefore the various sets of 22 glyphs commonly found in comparative tables of Semitic *abjads* (for example, those shown in Figure 1 below, and in <http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2311>, excluding the first two tables which show scripts with additional characters) are understood as sets of glyph variants of one another, and so of the one such script which has already been encoded in Unicode, the Hebrew script. (The Syriac script might also be considered a set of glyph variants, but its cursive joining behaviour and its modern use as a distinct script justify its separate encoding.) Further evidence for this position is that such comparative tables show continuous variation of glyphs rather than clearly distinct character sets; this is especially clear in Figure 1 below (see also the explanatory notes in Figure 2).

The strong negative reaction to the current proposal may well be because of a perception that separate encoding of Phoenician script would undermine the existing scholarly practice of replacing Phoenician glyphs with Hebrew ones by declaring it to be non-standard. There is also serious concern that in future, if new Phoenician characters are accepted, some texts which have been preserved with Phoenician glyphs will be represented with the new characters and others with the existing Unicode Hebrew characters. This is expected to result in confusion, impede scholarly work by complicating searches etc, and generally work against the Unicode goal of standardisation.

The need for a plain text distinction

The proposer has not demonstrated that anyone has a need, and not only a desire, to encode Phoenician letters as distinct plain text, rather than as graphics or as a marked up text using the abstract characters already encoded for Hebrew. Indeed, originally he claimed to have made no contact with the user community, although in the revised proposal he mentions indirect contact through Deborah Anderson. Only one of the named contacts in the revised proposal is in fact a scholar of Semitic languages, Jo Ann Hackett. She is cited in support of the proposal, but her edited comments do not demonstrate that she is aware of the basic Unicode design principle of encoding characters, not glyphs.

Others have suggested scenarios in which some users of Phoenician writing might need to make a plain text distinction. But these scenarios involve only occasional use of Phoenician glyphs by those who do not use them regularly. Therefore, if Unicode support is to be provided for them this should be done in a way which is acceptable to and does not conflict with the understanding and interests of the community of scholars who use these glyphs on a daily basis. Nevertheless, there is some evidence of a small scale requirement for making plain text distinctions between the different glyph variants of the same set of north-west Semitic abstract characters, for example between the Phoenician and square Aramaic/Hebrew variants.

Distinguishing between glyph variants in Unicode

The Unicode standard recognises that sometimes distinctions need to be made in plain text between different glyph variants of the same abstract characters. The following is taken from the Unicode Standard version 4.0.0, section 15.6, p.397:

Occasionally the need arises in text processing to restrict or change the set of glyphs that are to be used to represent a character. Normally such changes are indicated by choice of font or style in rich text documents. In special circumstances, such a variation from the normal range of appearance needs to be expressed side-by-side in the same document in plain text contexts, where it is impossible or inconvenient to exchange formatted text.

This seems to fit well with the occasional requirement to distinguish in plain text between the glyph variants of the same abstract characters in the north-west Semitic script. The mechanism defined in this section for making such distinctions is variation selectors. However, in response to a suggestion that Phoenician writing might be represented as Unicode Hebrew characters with variation selectors, Kenneth Whistler wrote on 2004-05-20 (<http://www.unicode.org/mail-arch/unicode-ml/y2004-m05/1140.html>): "the UTC has never had any intention that variation sequences be used this way -- and as a result would never acquiesce in encoding an entire **script** as a set of variation sequences off another script." But this response is based on his presupposition that Phoenician is a separate script.

An alternative mechanism which might be considered is to encode the Phoenician glyphs as separate characters but with compatibility decompositions to the existing Hebrew characters (the Unicode stability policy implies that this cannot be done *vice versa*). They are not compatibility characters in the defined sense that they "would not have been encoded except for compatibility and round-trip convertibility with other standards"; in this they resemble the Mathematical Alphanumeric Symbols which have compatibility decompositions. Nevertheless, these decompositions would indicate that the Phoenician characters "are variants of characters that already have encodings as normal (that is, non-compatibility) characters in the Unicode Standard" (quotations from the Unicode Standard version 4.0.0, section 2.3, p.23). However, Kenneth Whistler has also rejected this as a mechanism for making a plain text distinction between variants of the same set of abstract characters.

If in this case there is a need to make plain text distinctions between glyph variants which cannot be suitably met by the existing defined mechanisms, there may be a need to define a new mechanism. However, for implementation reasons this should be considered a last resort. It would be preferable to extend the scope of an existing mechanism which is already adequate or nearly so, such as the two mechanisms described above.

The name of the proposed characters

Additionally, the proposed name of the new characters is misleading. Phoenician is only one of a number of languages which were commonly written with the style of glyph used in the proposal. There is a close analogy with the Old Italic script, which, although commonly called "Etruscan", was named "Old Italic" for Unicode because it is used for several languages and not just Etruscan. To avoid further confusion between language, script and glyphs, the Unicode name for these glyphs should be not "Phoenician" but a more generic term. The suggested name "Old Canaanite" is a good one.

Recommendations to the UTC

This response can be summarised as two sets of recommendation to the UTC, as follows:

Recommendations concerning criteria for unifying historic scripts

Concerning the review of factors that should be taken into account in determining whether to unify two scripts or not, the following is recommended to the UTC:

1. The review should be completed before a decision is taken on any specific proposal for a new historic script, including the Phoenician script proposal.
2. The review should uphold the basic principle that Unicode encodes characters, not glyphs, by requiring that proposals for new historic scripts should demonstrate that the proposed script is indeed a separate set of abstract characters, and not a set of glyph variants of already encoded characters.
3. The principle should be established that decisions on whether any proposed script is a separate set of characters should be taken in close consultation with leading scholarly experts on the specific form of writing.
4. Issues of utility and user requests for plain text distinctions should be considered secondary and not allowed to confuse discussions on the principle of whether the proposed script is a separate set of characters.
5. There should be no presumption that the existing Roadmap is either theoretically correct or represents the most advisable set of scripts to be encoded.

Recommendations concerning the Phoenician script proposal

Concerning the Phoenician script proposal, it is recommended that the UTC, when considering this proposal, should look at the following questions in order:

1. Are the proposed Phoenician characters in fact distinct abstract characters (the proposer's view), or are they glyph variants of characters which are already encoded in the Unicode Hebrew block (the view of the community of scholars of these forms of writing)? If the proposer's view is accepted, the proposal may then be accepted without further question (except about the names of the characters), but a consequence will be a serious loss of

credibility of the Unicode standard among the scholarly community, as well as confusion among users.

2. If Phoenician letters are not distinct abstract characters, is there in practice an exceptional need to distinguish them *in plain text* from the existing Hebrew characters, a need which cannot be met adequately by use of mark-up and font distinctions, and which is sufficiently non-trivial that the Private Use Area is not appropriate? If this cannot be demonstrated, the proposal should simply be rejected.
3. If there is a need to make a plain text distinction between glyph variants, what is the appropriate mechanism for making this distinction? Possible mechanisms include encoding the proposed Phoenician characters as variation sequences or as separate characters with compatibility decompositions. UTC members may wish to suggest other suitable mechanisms.

These questions, and especially the last one, may be sufficiently complex to justify deferment of this proposal while UTC members take further expert advice.

NOTE ON TABLE OF ALPHABETS

1. The Inscription of Mesha, see § 2 *d*.
2. Old Hebrew seals and gems, see § 2 *d*.
3. Siloam inscription, see § 2 *d*.
- 4-6. Jewish Coins, see § 2 *d*.
7. From Lidzbarski, *Nordsemitische Epigraphik*, plate xxi, 7, 8; and Ph. Berger, *Histoire de l'écriture*, ed. 2, p. 200.
8. From the Palaeographical Society's *Oriental Series*, plate xxviii.
9. From Ad. Neubauer, *Facsimiles of Hebr. MSS.*, Oxford, 1886, pl. xl.
10. Inscriptions of Zinjīth, see Lidzbarski, *ibid.* i, p. 440 ff., and plates xxii-xxiv; and Cooke, *Handbook of North-Semitic Inscriptions*.
11. Old Aramaic inscriptions on weights, seals, vessels, and tablets. See [Georgus] *Inscriptionum Semiticarum*, ii, pl. i ff., and Cooke, *ibid.*
12. Inscriptions of Teima, *CIS*, ii, pl. ix f.; and Cooke, *ibid.*
13. Papyrus and Ostraka from Egypt, *CIS*, ii, pl. xii ff.; and Cooke, *ibid.*
14. Early Hebrew inscriptions in square character, from Lidzbarski, *ibid.*, pl. xlii; and *Beitriige zur Assyriologie*, iv, plate 10 p. 203.
15. From the facsimile of the Codex Babylonicus Petropolitanus (916 A.D.) in Chwolson, *Coryus Inser. Hebr.*, no. 103.
16. From Neubauer, *Facsimiles*, pl. viii.
17. From the Palaeographical Society, pl. xli, completed from Neubauer, pl. xiv.
18. From a modern MS. on parchment.
19. From Mrs. Lewis and Mrs. Gibson, *Palestinae Syriae Texts*, London, 1900, pl. iii.
20. From Neubauer, pl. ix. (Most like the printed 'Rashi' character.)
21. From the Palaeographical Society, pl. xli, completed from Neubauer, pl. xiv.
22. From Neubauer, plates xix and xx.
23. From Kauting, *Table of Semitic Characters*, col. 49.
24. From Neubauer, pl. x.
25. From Neubauer, plates xxviii and xxx.
26. From a modern MS.

Figure 2: Note on the Table of Alphabets, Figure 1.