

Title: A 3-tier system for Ancient Egyptian hieroglyphs.

From: Mark-Jan Nederhof, Peter Dils, Andrew Glass, Jorke Grotenhuis, Svenja Gülden, Stéphane Polis, Daniel A. Werning

Date: 2024-06-02

Abstract

Previous documents proposed a 2-tier system for the repertoire of Ancient Egyptian hieroglyphs. Here we propose moving to a 3-tier system. The motivation is to avoid potential inconsistencies between applications and fonts that could undermine users’ trust in Unicode for representing hieroglyphic text.

This document refers to:

- Michel Suignard et al. (2023), *Encoding proposal for an extended Egyptian Hieroglyphs repertoire*, L2/23-181R2
- Daniel A. Werning et al. (2024), *Additional variation selectors for rotations of Ancient Egyptian hieroglyphic texts*, proposal currently under consideration

1 Context

1.1 Extended signlist

The recently accepted proposal by Suignard et al. includes almost 4,000 new hieroglyphs. Among these are **core** signs that are verified and uncontroversial, and **non-core** signs that could not yet be verified or whose use for some other reason cannot be endorsed by domain experts as yet. Users would generally be advised to restrict themselves to core signs if possible, and avoid non-core signs for the time being. Cf.:






The “Core list” is a curated subset of characters from the FULL LIST. It is the recommended set for Egyptologists and should be implemented in widely used fonts. (Op. cit. p. 4)

Implementation of hieroglyphic fonts is complicated by the inability of modern font technology to dynamically scale glyphs. Consequently, multiple sizes of each glyph need to be stored in a font. During rendering of hieroglyphic text, the appropriate size of each sign occurrence is chosen depending on possible use of control characters, such as horizontal and vertical joiners, and on dimensions of neighboring signs. Due to the costs of storing multiple sizes of many hundreds of non-core signs, most of which would never be used, it may be advisable for font designers to forego implementation of non-core signs, or to not implement them fully, that is, they would typically not interact with control characters.

1.2 Rotation

Some signs occasionally occur in a rotated form. Three variation selectors can be used for rotation by multiples of 90°, and four further variation selectors have been reserved for other angles of rotation. An updated list of variation sequences has recently been submitted by Werning et al. involving all seven variation selectors.

A consensus was reached that it is preferable for a rotated sign to be henceforth represented as the combination of a base sign and a variation selector, even if that rotated form already existed as a single code point.

For example, among ,  and , the diagonal orientation  is the more typical of the three. If  = U+13338 is chosen as the ‘base’ sign, the other two orientations can be encoded as:

$$\begin{array}{l}
 \text{A} \leftarrow \text{A} \left[\begin{array}{c} \overline{\text{VS}} \\ 4 \end{array} \right] = \text{U+13338 U+FE03} \\
 \text{A} \leftarrow \text{A} \left[\begin{array}{c} \overline{\text{VS}} \\ 7 \end{array} \right] = \text{U+13338 U+FE06}
 \end{array}$$

However, A and A also exist as single code points U+13339 and U+1333B , respectively. As existing code points cannot be removed from Unicode, the variation sequences create encoding ambiguity whereby, for example, the grapheme A can be encoded either as U+1333B or as U+13338 U+FE06 .

2 Problem

Previously it was assumed that code points like $\text{A} = \text{U+1333B}$ would be declared non-core, to discourage encoders from using them and prescribe use of newly introduced variation sequences like $\text{A} \left[\begin{array}{c} \overline{\text{VS}} \\ 7 \end{array} \right]$ instead. This means that there would be at least two kinds of non-core signs:

- newly introduced signs that have not been verified, and
- rotated signs, whose code points were widely used in the past, but that should no longer be used, to avoid encoding ambiguity.

In an idealized scenario, code points of rotated signs will eventually disappear, which thereby makes the encoding ambiguity moot. At that stage, there would no longer be any need for fonts to support such code points. However, this scenario may not materialize in the foreseeable future or possibly ever. The reason is that as long as a few popular fonts support such code points, there may not be any compelling reason for many encoders to avoid using them; that such signs happen to be listed as ‘non-core’ on the Unicode web site may not be part of their considerations.

Because such signs are declared non-core, some font designers however may well choose to omit them from their fonts, or not implement them fully, just as they would not implement some newly introduced, unverified non-core signs. This could lead to a situation where some tools and some fonts implement such code points and others do not. Given that A and several other such rotated signs are extremely frequent, there is a high probability that users will witness inconsistent behavior when they access their encodings using different tools and different fonts. This could undermine their trust in Unicode as a standard, which is normally expected to exhibit consistent behavior.

3 Solution

The described problem is caused by conflating two kinds of signs. The most obvious solution is therefore to distinguish them, resulting altogether in the following 3-tier system:

Core signs have been verified and can be used without reservations.

Non-core signs have not been verified and they are best not used unless there are compelling reasons to.

Legacy signs had legitimate uses in the past, but are best not used anymore, because of changes in encoding practices.

Code points may migrate between these three tiers, but subject to a number of constraints. Non-core signs may become core once they have been verified. But if a code point has ever been core, it can only become legacy. This could happen if an alternative encoding of the same grapheme results from introduction of a

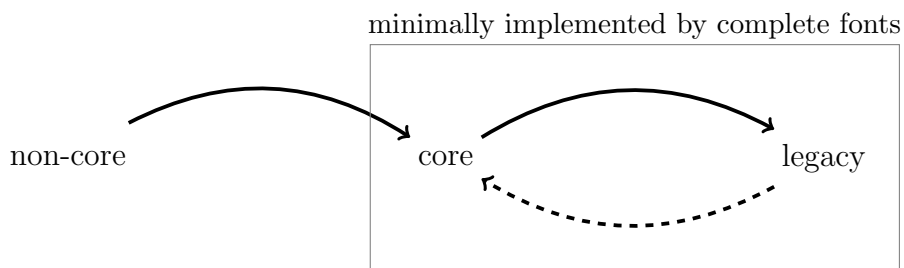


Figure 1: Potential migration of signs within the 3-tier system

new control character or variation sequence, making the code point redundant. It could also happen that our understanding of the status of a shape as independent grapheme changes, in which case a sign may migrate from core to legacy, and conceivably from legacy back to core (Figure 1).

A complete font for Ancient Egyptian hieroglyphs is expected to implement both core and legacy signs. But fonts are not required to fully implement all non-core signs. As a consequence of the constraints on migration between the three tiers, a user can be confident that any complete font will correctly render an encoding, as long as the encoding does not contain non-core code points, and this will remain so for any given encoding, regardless of future developments in Unicode, such as addition of new hieroglyphs to the sign list or addition of new control characters.

4 Properties

For the 2-tier system, the `kEH_Core` value would have been either `Y` or `N`. If transition to the proposed 3-tier system is accepted, then the value of `kEH_Core` would be one of:

`Y` for core signs

`N` for non-core signs

`L` for legacy signs

5 Conclusions

The proposed 3-tier system may not appear to be drastically different from the 2-tier system proposed earlier. As before, there are signs that can be used without reservations, and there are signs that are best not used. Carefully curated Egyptological databases may exclude both non-core and legacy signs, and thereby the distinction between the 2-tier and 3-tier system is largely inconsequential for specialist users.

However, for everyday use, by professional Egyptologists as well as enthusiasts, who may wish to copy existing encodings from different sources and transfer encodings between applications, the distinction between legacy signs and (other) non-core signs may well become relevant. The concepts of ‘core’ and ‘non-core’ as part of the 2-tier system are relatively new, and it is not certain how font designers would differentiate implementation of existing and new signs depending on the core/non-core distinction. The introduction of the additional concept of ‘legacy’ aims to remove much of this uncertainty. By the explicit advice to font designers to implement both core and legacy signs, the 3-tier system thereby provides better guarantees for a consistent user experience.