

Title: Comments on three control characters for Egyptian Hieroglyphs
 From: Mark-Jan Nederhof and Vinodh Rajan (University of St Andrews, UK)
 To: UTC
 Date: 2016-04-18

1 Introduction

This document comments on **L2/16-018** by Bob Richmond and Andrew Glass, which proposes three control characters:

- EGYPTIAN HIEROGLYPHIC SIGN LIGATURE JOINER
- EGYPTIAN HIEROGLYPHIC SIGN HORIZONTAL JOINER
- EGYPTIAN HIEROGLYPHIC SIGN VERTICAL JOINER

We will argue that the first of these, the EGYPTIAN HIEROGLYPHIC SIGN LIGATURE JOINER, conflicts with the fundamental principles of Unicode, as the meaning of that character is inherently **undefined**. Its use will cause significant errors to be introduced into scholarly work.

2 Background

Ancient Egyptian hieroglyphic writing sets itself apart from most other writing systems in the world in at least four respects:

1. The repertoire of signs (hieroglyphs) was never stable. New signs could be introduced with relative ease, while other signs fell out of use.
2. There was a flexible boundary between the meaning of what was depicted by a sign and how it was depicted (cf. 'glyphs') and the more abstract purposes of signs to form parts of words without reference to their exact appearance (cf. 'characters'). This causes further complications for precisely defining the sign repertoire [3].
3. Words could be written in many different ways, even within the same text.
4. Aesthetics of the spatial layout, in particular avoidance of ugly white-space, was of the utmost importance in monumental inscriptions.



The first and second of these issues explain why it took so long to include the 1071 hieroglyphs currently in Unicode (from version 5.2), which is a small subset of all hieroglyphs used throughout the history of Ancient Egypt, and it explains why the Egyptological community has, as yet, been unable to agree on a coherent strategy how to compile a more comprehensive list of signs in a principled way, despite well-argued attempts in this direction [9].¹

The third and fourth of the above issues are responsible for the difficulty of encoding the spatial arrangement of signs. One may observe that two wide flat signs were typically placed one

¹Problematic are documents such as L2/16-028 by Michel Suignard, which indiscriminately bring up thousands upon thousands of undocumented shapes, without providing any suggestion how to even start the necessary analysis of determining where signs were found (possibly as hapax), what signs mean, whether they are independent characters or graphical variants, and overall what their justification would be for occupying a code point on their own.

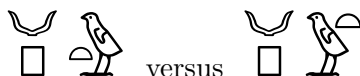
strictly above the other (without the bounding boxes of the two signs overlapping). Similarly, two tall narrow signs were typically placed one strictly beside the other. We will refer to these two examples as 'orthogonal' arrangements of signs. One may also frequently observe that a small sign was placed in the empty corner of the bounding box of another sign (or in a virtual box slightly larger than the bounding box). We will call this a 'non-orthogonal' arrangement of signs.

The aesthetical quality of writing in Ancient Egypt was so important that it sometimes motivated swapping the order of signs, relative to the reading order [5, p. 51]. We illustrate such


'graphic transposition' by the two non-orthogonal groups  and . Here the chick stands for (the sound commonly transliterated as) **w** and the half-circle-shaped sign ('bread') stands for **t**. The normal left-to-right reading order would therefore suggest the readings **tw** and **wt**, respectively. Both **tw** and **wt** correspond to common morphemes. The first can, among other things, be a marker for passive and a demonstrative pronoun. The second can, among other things, be a phonetic writing of the feminine plural ending, and the root of a verb meaning 'to bandage'. Most commonly **tw** is written as the first group above and **wt** is written as the second group above, as one would expect, but the first group also frequently stands for **wt**.

However, it is wrong to say that the two groups are exchangeable. They have a different distribution, and whether the first group more likely stands for **tw** or for **wt** depends on the context in the writing of a word, on the period, and on the genre of text.

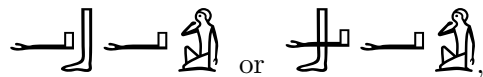
As an example, the following are conceivable writings of **wp.tw**, 'is opened', and **wpwt**, 'message', respectively [4, volume 1, pp. 299-303]:



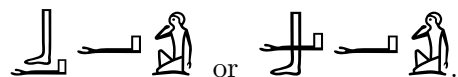
(Here the first two signs together account for **wp**.)

Something similar holds for superimposed signs, as for example . Here the leg stands for **b** and the arm stands for the 'ayin', which we will pragmatically transliterate here as **a**. The sign combination as a whole commonly reads **ab**, but it may also read **ba** [4, volume 1, pp. 173-179 and pp. 446-450]. In the first case, the group may also be written as detached signs for **a** and **b**, in this order. In the second case, the group may also be written as detached signs for **b** and **a**, in this order. Therefore the superimposed group is not exchangeable with a detached group, in either order.

As an example, conceivable writings of **aba**, 'boast' [4, volume 1, p. 177] could be



while conceivable writings of **baa**, 'drink blood' [4, volume 1, pp. 446-447] could be



(Here the sitting man with his hand to his mouth is a determinative.) The second writings of these two words happen to be the same, but the first writings are not exchangeable. Note all three writings are to some extent non-orthogonal, in the sense that bounding boxes of signs overlap, and are thereby relevant to the discussion below.

It follows that for Ancient Egyptian text it is essential that we do not confuse:


- the order in which signs are written,
- whether signs are superimposed or not.

3 The '&' operator

Among several encoding schemes for hieroglyphic text proposed in the 1980s, the one that has become most wide-spread is the Manuel de Codage (MdC). The main elements for expressing spatial arrangements of signs were the operators ':', to indicate one sign (or group) is above another, and '* ', to indicate one sign (or group) is next to another. The '* ' has higher operator precedence than ':'. Round brackets can be used to override operator precedence. The ':', with operator precedence lower than those of '* ' and ':', denotes horizontal arrangement of groups of signs for horizontal text (i.e. written in rows) and vertical arrangement for vertical text (i.e. written in columns).

The origins of the MdC are intimately tied to the development of Glyph, which was a commercial software product [1]. It was meant to facilitate **printed** publications of hieroglyphic texts. As such, the developers had no qualms to include new features into the MdC that only had meanings within a certain version of this particular software product, without any intention to formally document them.

One such feature was the '&' operator, which is not part of any formal standard, although it occurs in examples (without any discussion or explanation) in an **unfinished** electronic document by one of the later developers of Glyph [12]. The '&' operator was used to combine two or more signs in **any** non-orthogonal arrangement. For example, G39&N5 would denote sign G39 ('pintail'), with N5 ('sun') inserted in the upper right corner of the bounding box of G39, and

scaled down to fit: .

That G39&N5 means this, rather than say N5 inserted in the lower left corner of the bounding box of G39, is fixed and defined by the tool. However, users could themselves extend the meaning of '&' to apply to combinations of signs beyond those defined by the tool. This last possibility is essential because any hieroglyphic text one investigates is likely to have many non-orthogonal arrangements of signs that one has never seen before.






The reason this was not a problem at the time is that the MdC was not intended for standardization, nor for exchange of encodings between scholars, but only for preparing **printed** publications. After publication, the encodings could be safely discarded. Compatibility between implementations of MdC was a non-issue. However, bringing the '&' into the world of Unicode, as the EGYPTIAN HIEROGLYPHIC SIGN LIGATURE JOINER, is a grave mistake, as Unicode does aim to achieve compatibility and standardization.

We will illustrate the problems using the EGPZ 1.0 BETA Specification [10], which lists sign combinations using '&'. There is no doubt this document contains only a tiny fraction out of an inexhaustible supply of non-orthogonal sign arrangements one may find in hieroglyphic texts. We will show that the meaning of '&' is unpredictable, and thereby not just ill-defined, but undefined.








Suppose sign A depicts a bird, then is B in A&B placed in front of A, or in the upper right corner or in the lower right corner?






U+eb6d G29&N29

	U+eb89	G43&O4
	U+eb61	G17&V13
	U+eb86	G43&N21
	U+eb6e	G29&R7
	U+eb6c	G28&X1
	U+eb7e	G39&X1&N5

Suppose for sign A, there is empty space in the lower left corner and empty space in the lower right corner, then where is B placed in A&B ?

	U+eb2a	F29&X1
	U+eb6c	G28&X1
	U+eb6e	G29&R7
	U+eb6b	G26&X1&Z4
	U+ebf5	O44&X1&Z1
	U+ebfa	P6&D36&N5&Z1
	U+eb29	F28&T11&X1&X1

Suppose for sign A, there is empty space in one or more corners, and suppose sign B is three strokes (plural determinative), then where is B placed in A&B ?

	U+eb00	A14&Z2
	U+eb01	A17&Z2
	U+eb02	A24&Z2d

Suppose for sign A, there is empty space in the upper right corner and empty space below it, and B is a single stroke (semogram marker), then where is B placed in A&B ?



U+eb15 E6&Z1



U+eb18 E10&Z1

Suppose for sign A, there is empty space in the upper right corner, and empty space elsewhere, and suppose B and C are small signs, then where are B and C placed in A&B&C ?



U+eb05 D17&N5&X1



U+ec21 U21&N5&Z1



U+eb40 G1&X1&W11



U+eb31 F44&N5&X1



U+ec57 X1&G39&X1&Z1&X1



U+ec59 X1&G39&Z1&X1&X1



U+ec22 U21&Q3&Y1

For wide and flat sign B, should B be below A in A&B, or should the two signs be superimposed?



U+eb09 D28&D52



U+ec3f W14&I9



U+ebf0 N104&D40

Suppose A is a sign such as the cobra and B is a tall sign, should B be strictly below A or should the signs be superimposed, in A&B ?



U+eba3 I10&D58



U+ebad I10&M13

In some of these examples, Egyptologists might recognize common patterns of signs and might correctly predict where one sign is placed relative to another. But the proposed control character can be applied on any sign combination, including those that only occur in a single obscure inscription.

A concrete scenario is the following. One scholar encodes sign A to the **left** of sign B as A&B, because this is how A&B looks in their rendering engine, with A, '&' and B replaced by the corresponding Unicode characters. This scholar sends the encoding to a second scholar, whose rendering engine happens to print A&B as A to the **right** of B, and thus the text is wrongly rendered, say in PDF or in a printed publication. We can find a similar scenario that concerns superimposed or detached signs.

As we explained in Section 2, in Ancient Egyptian, the order of hieroglyphs in a text matters, and so does the question whether signs are superimposed or not. The mistakes that arise from the sketched scenarios are therefore significant, as opposed to mere typographical imperfections.

4 Conclusions

It is not uncommon that rendering engines do not know how to render a certain control character in combination with other characters. If they then indicate that they are unable to render the sequence of characters, using a default glyph for the control character, then there is little harm, as the information is still visible, and the meaning of the sequence of characters can be obtained by consulting Unicode documentation.

What we are dealing with here, in the case of the EGYPTIAN HIEROGLYPHIC SIGN LIGATURE JOINER in proposal L2/16-018, is very different. No exhaustive list of all combinations of signs joinable by this control character could ever be compiled and centrally maintained. Hence the meanings of these sign combinations would be defined by individual rendering engines, rather than by the Unicode standard. The scholarly errors that will inevitably result stem from the differences between these individual rendering engines.

We have no great objection to the other two control characters from L2/16-018, viz. the EGYPTIAN HIEROGLYPHIC SIGN HORIZONTAL JOINER (the equivalent of MdC '**') and the EGYPTIAN HIEROGLYPHIC SIGN VERTICAL JOINER (the equivalent of MdC ':'). We would contend however that these two signs alone are too limited for rendering hieroglyphic text in a plausible way. (There is also a problem with operator precedence; see Appendix A.)

Proposal L2/16-018 is correct in stating that without control characters, Unicode has limited usefulness in Egyptology, as a linearized sequence of hieroglyphs is not an acceptable representation of hieroglyphic text for Egyptologists. But introducing control characters whose meaning is undefined is not the solution. We further disagree with document L2/16-037, by Deborah Anderson et al.:

With the encoding of these three characters, quadrats can be created in plain text, and Egyptologists will no longer need to rely on proprietary software for rendering [...]

As Unicode leaves the meaning of the EGYPTIAN HIEROGLYPHIC SIGN LIGATURE JOINER undefined, it is unclear at this time which tool(s) will define it, if at all possible, and whether these tools will be proprietary. In contrast, the currently most popular tool to typeset hieroglyphic texts, JSesh [11], is not proprietary. Furthermore, the intricacies of typesetting hieroglyphic text, and the demands of Egyptologists for accurate rendering, should not

be underestimated. It is not likely that the need for dedicated hieroglyphic editors will disappear with adoption of the three, or any small number of, control characters in Unicode.

Our recommendation is to suspend adoption of all three control characters from L2/16-018 until a more principled and coherent scheme can be worked out, with control characters that can be assigned proper meanings. This could be modelled on a subset of RES [8], which has primitives for e.g. 'put one sign (or group) in the empty upper right corner of another sign (or group)' and 'superimpose two signs (or groups)'. The technical realization of control characters corresponding to these primitives should be no harder than that of '&', with the added benefit that the intended meaning is always fixed, regardless of whether a rendering engine knows how to realise the spatial layout of a control character for a particular choice of signs.

It should be pointed out that there are some existing characters in Unicode that correspond to the control characters that we suggest above, as the Ideographic Description Characters in the range U+2FF0 — U+2FFF, intended for CJK ideographs. At the very least, this constitutes a precedent for control characters that define spatial arrangements of characters.

Acknowledgments

This document draws upon fruitful correspondence with Serge Rosmorduc, Stéphane Polis and Simon Schweitzer.

References

- [1] J. Buurman, N. Grimal, M. Hainsworth, J. Hallof, and D. van der Plas. *Inventaire des signes hiéroglyphiques en vue de leur saisie informatique*. Institut de France, Paris, 1988.
- [2] M. Collier and B. Manley. *How to read Egyptian hieroglyphs: A step-by-step guide to teach yourself*. British Museum Press, 1998.
- [3] P. Collombert. Combien y avait-il de hiéroglyphes? *Egypte, Afrique et Orient*, 46:35–48, 2007.
- [4] A. Erman and H. Grapow. *Wörterbuch der Ägyptischen Sprache*. Akademie-Verlag, Berlin, 1926–1961.
- [5] A. Gardiner. *Egyptian Grammar*. Griffith Institute, Ashmolean Museum, Oxford, 1957.
- [6] M.-J. Nederhof. A revised encoding scheme for hieroglyphic. In *Proceedings of the 14th Table Ronde Informatique et Égyptologie*, July 2002. On CD-ROM.
- [7] M.-J. Nederhof. The Manuel de Codage encoding of hieroglyphs impedes development of corpora. In S. Polis and J. Winand, editors, *Texts, Languages & Information Technology in Egyptology*, pages 103–110. Presses Universitaires de Liège, 2013.
- [8] M.-J. Nederhof. RES (Revised Encoding Scheme). <http://mjn.host.cs.st-andrews.ac.uk/egyptian/res/>, 2016.
- [9] S. Polis and S. Rosmorduc. Réviser le codage de l'égyptien ancien. Vers un répertoire partagé des signes hiéroglyphiques. *Document Numérique*, 16(3):45–67, 2013.
- [10] B. Richmond. EGPZ 1.0 beta specification. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.182.4847>, 2007.

- [11] S. Rosmorduc. JSesh Hieroglyphic Editor. <http://jseshdoc.qenherkhopeshef.org>, 2016.
- [12] H. van den Berg. Manuel de Codage: A standard system for the computer-encoding of Egyptian transliteration and hieroglyphic texts. <http://www.catchpenny.org/codage/>, 1997. Accessed 2016-04-05.


A Operator precedence

We explained at the beginning of Section 3 that operator precedence in the MdC can be overridden using brackets. Document L2/16-018 claims:


An extensive survey of the Egyptian texts indicates there is no need to support parenthetical expressions for quadrats, since parenthetical expressions occurring in Mdd can be expressed as sequences using the LIGATURE JOINER character or other mechanisms [...]

This is most certainly false. A superficial inspection of just a few common, well-known texts taken from a widely available publication [2] reveals that restriction to ':', '*' and '-' with the default operator precedence is inadequate for horizontal text (the problem is even greater for vertical text).


For example, in BM EA 581 [2, p. 59] we find:

 N35:(W24:X1)*Z1*Z1


In BM EA 1873 [2, p. 74] we find:

 N35:N35:(W24:X1)*G43

and:

 N35:V28*(N29:D21)

In BM EA 571 [2, p. 79] we find:

 N41*(X1:U7):I9

To reiterate a point made before [6, 7], it is a mistake to try to judge the adequacy of an encoding scheme for **original** hieroglyphic text by only comparing it against **typeset** hieroglyphic text that was previously produced using outdated technology suffering from severe limitations. Where the considered typeset hieroglyphic was itself produced using some variant of the MdC, the reasoning in favour of adequacy of elements from the MdC becomes circular and vacuous.





18. April 2016


Remark on the ‘Proposal to encode three control characters for Egyptian Hieroglyphs’ (L2/16-018R) by Bob Richmond and Andrew Glass and on ‘Comments on three control characters for Egyptian Hieroglyphs’ by Mark-Jan Nederhof and Vinodh Rajan


The integration of the Egyptian writing system into Unicode would be an enormous achievement for Egyptian corpus linguistics and as such is an indispensable task, although a difficult one. The undersigned are writing on behalf of the *Thesaurus Linguae Aegyptiae* <<http://aaew.bbaw.de/tla>> (TLA), with 1.4 million tokens the largest electronic corpus of Egyptian full text data.

We highly appreciate the ongoing Unicode initiative, including the attempt to establish control characters to render hieroglyphic quadrats by Bob Richmond and Andrew Glass (L2/16-018R), and we fully agree with Richmond’s and Glass’s suggestion to have the two characters EGYPTIAN HIEROGLYPH HORIZONTAL JOINER and EGYPTIAN HIEROGLYPH VERTICAL JOINER. The proposed EGYPTIAN HIEROGLYPH LIGATURE JOINER, however, would apparently generate inconsistent results. There should be an

opportunity, for example, to encode two different quadrats #1  and #2 , thus the characters U+1315C, U+133CF, U+1340D in combination with the proposed U+13430. But there is no definition of the sequence of the characters, and a rendering engine cannot make a decision whether to group U+1315C U+13430 U+133CF U+13430 U+1340D into quadrat #1 or quadrat #2. Therefore we share Mark-Jan Nederhof’s and Vinodh Rajan’s concern about the introduction of an EGYPTIAN HIEROGLYPH LIGATURE JOINER since this would lead to ambiguous encodings of Egyptian hieroglyphs.

Richmond and Glass argued that “LIGATURE JOINER is the highest priority in the order of precedence of Egyptian Joiners.” Nederhof and Rajan already collected some examples contradicting this opinion. We underline the validity of these examples and want to point to another special case related to the issue of the order of precedence: the case of ligatures

with hieroglyphs grouped horizontally and vertically, as in the square  consisting of the characters U+13113, U+1340D, U+133CF, U+1341D. As mentioned above, the encoding X+X+X+X alone cannot guarantee the exact grouping of the hieroglyphs. The encoding

X+X*X:X would however wrongly produce . So we still seem to be in need of brackets to specify the order of precedence.

In one word, the integration of EGYPTIAN HIEROGLYPH HORIZONTAL JOINER and EGYPTIAN HIEROGLYPH VERTICAL JOINER is perfectly right and desirable, whereas EGYPTIAN HIEROGLYPH LIGATURE JOINER would result in confusing outcomes and is not suitable for Egyptian corpus linguistics. We do need characters to encode operator precedence and positioning of hieroglyphs in ligatures.

Prof. Dr. Tonio Sebastian Richter
TLA, Project Director

Dr. Ingelore Hafemann
TLA, Research Coordinator

Dr. Simon Schweitzer
TLA, Researcher




Notes about the ‘Proposal to encode three control characters for Egyptian Hieroglyphs’ (L2/16-018R) by Bob Richmond and Andrew Glass

The undersigned are writing on behalf of the “Ramses Project” and *Ramses Online* (ramses.ulg.ac.be), a richly annotated corpus of Late Egyptian texts. The 530.000 tokens of this corpus are lemmatized, annotated for PoS and inflexion, and — crucially in this context — the hieroglyphic spellings of all the words are encoded (68 000 different spellings).


We would first like to stress that the inclusion of the hieroglyphic writing system into Unicode is of paramount importance for the future of our field and we highly appreciate all the efforts made in this direction. We also agree with Richmond and Glass’ observation in the introduction of their “Proposal to encode three control characters for Egyptian hieroglyphs”: the fact that Egyptian hieroglyphs cannot be displayed in plain text using the quadratic format, a feature integral to the writing system, has resulted in a very limited use of Unicode for hieroglyphic text, which is a pity.

It should be noticed that hieroglyphic text composition is currently used in a number of contexts : short quotations of words or sentences in the middle of scientific papers or books, normalized transcription of an original text in a scientific edition, and, more recently, uses in databases, like the *TLA* or *Ramses Online*.

In the proposal L2/16/018, two operators (HORIZONTAL and VERTICAL JOINER) are reasonably well defined, and, if supported by software solutions, will be useful for tasks where the fidelity to the original document is not a prime requirement, or where a limited rendering is better than no rendering at all (inclusion in databases like access or framemaker, for instance). We do support their addition.

The EGYPTIAN HIEROGLYPHIC SIGN LIGATURE JOINER, however, is misleading and dangerous in unicode, because its effect is not well defined. In this respect, we heartfully agree with the observation made by Mark-Jan Nederhof and Vinodh Rajan in their ‘Comments on the three control characters for Egyptian Hieroglyphs’ and think that a more coherent scheme should be worked out. Taking a simple example from our database, the sequence EGYPTIAN HIEROGLYPH G035 EGYPTIAN HIEROGLYPH X007 can be composed as ,  or . Even though there is a statistical bias toward the first grouping, the second is quite frequent too and the last one is the norm in hieroglyphic transcriptions of hieratic

sources. The risk is that we would have either a well defined (but hard to extend) list of frequent groupings, which would be secure but would contradict somehow the unicode philosophy, or some automated mechanism, which might result in different rendering across platforms. Given the status of Unicode as a lasting standard, we think that it is more appropriate to avoid adding such an operator in it.

Furthermore, as noted by Mark-Jan Nederhof and Vinodh Rajan, the highest priority in the order of precedence attributed to the EGYPTIAN HIEROGLYPHIC SIGN LIGATURE JOINER does not solve all the problems in terms of positioning of signs.  (Stela Cairo, JE 60539, l. 8), for instance, cannot be handled staisfactorily with the present proposal : we still need some bracketing system in order to specify the order of precedence or several control characters (if bracketing is not an option).

Our best wishes,

Stéphane POLIS
Research Associate FRS-F.N.R.S.

Serge ROSMORDUC
CNAM, Paris