The Manuel de Codage encoding of hieroglyphs impedes development of corpora*

Mark-Jan NEDERHOF

University of St Andrews

1. INTRODUCTION

The Ancient Egyptian hieroglyphic writing system has a number of properties that set it apart from most other modern and ancient writing systems (Daniels & Bright 1996). One is that the pictographic aspect was maintained throughout its history. Stylisation and abbreviation of signs have played a much smaller role than in the cases of for example Akkadian cuneiform or Chinese. Whereas hieratic can be seen as a cursive form of hieroglyphic, the latter was never replaced by the former, and they influenced one another throughout history. Despite this moderate degree of character stylisation, there was no limit on the number of signs that could be used, and large variation can be observed in their exact appearance.

A second aspect of hieroglyphic writing that sets it apart is a particular form of aesthetics, including a desire to divide the available surface in a way pleasing to the eye, avoiding large empty spaces. Thus, two signs with large height and small width could be placed one next to the other, and two signs with small height and large width could be placed one below the other. This is however by no means the only way of placing signs relative to one another. Frequently, the empty space in the corner of one sign is used to harbour a second sign of small size. One sign can also be placed inside another or two signs can be positioned one overlapping the other, especially if there is a linguistic connection (e.g. collocation) between the words the two signs represent.

Because Ancient Egyptian hieroglyphs seem to form an exceptional case within the wide range of the world's writing systems, it is not *a priori* clear that common technical solutions that have been devised for processing other writing systems are also suitable to hieroglyphs. A good illustration is perhaps the arduous process that has led to the inclusion of hieroglyphs in Unicode.

The first proposal to include Ancient Egyptian was proposal N1637, undertaken by Michael Everson (1997). This was based on the sign list from Gardiner (^{3}1957) and comprised 761 signs, together with operators to encode relative positioning of signs, as found in the Manuel de Codage (see §2 below for further discussion of the Manuel de Codage). With proposal N1944 (Everson 1999a), this was extended to several thousands of signs, incorporating signs from Grimal *et al.* (1993). Both proposals drew quite some criticism, for example from Wolfgang Schenkel (Schenkel 1999; Everson 1999b).

The third attempt was more modest, with a total of 1071 signs, including the signs from Gardiner (1957) plus those from its supplements (Gardiner 1928, 1929, 1931, 1953) and a few signs from other

* I gratefully acknowledge fruitful discussions with Serge Rosmorduc. I am also very grateful to Horst Beinlich and Norbert Stief for correspondence about PLOTTEXT. Many thanks go to anonymous reviewers for a large number of improvements to the text.

sources, but leaving out the formatting operators. This list was finally approved and added to Unicode 5.2 as part of Plane 1 (Unicode 2010). (Plane 1 is a range of code points that is mostly used for characters from historic scripts.)

One factor that marred the discussions leading up to the eventual Unicode set of hieroglyphs was the disparity between the formal notion of 'character' and standard practices in Egyptology when transcribing hieratic or normalising hieroglyphic inscriptions. Following the terminology of Unicode, a *character* is the smallest component of written language and a *glyph* is a shape that a character can have when it is rendered or displayed. In Egyptology however, there seem to be tendencies to remain true to the original manuscript while encoding a text, often to the extent of encoding glyphs rather than characters.

An example is the distinction between  (G43) and  (Z7), which could be argued to be different shapes representing the same character. One also sees occurrences of  (W17) next to  (W18), which are different glyphs for the same character.

The cause of much of the confusion is the sign list by Gardiner, or perhaps more accurately put, its current misinterpretation. The intention of this list was never to create a list of characters in the sense of Unicode, but firstly to offer students an overview of different hieroglyphs and their functions and meanings, and secondly, to create an inventory of signs needed to print texts. In particular, for some signs, there is more than one glyph to be used by the printer in different contexts, such as G36a next to  G36 and  (N25a) next to (N25).

The problem of what information to represent in an encoding of hieroglyphic text not only pertains to the sign list but also involves the formatting, or in other words, how signs are positioned relative to one another. It is not always clear to what extent this aspect is important to encoding: on the one hand the relative positions of signs have little linguistic significance, whereas on the other hand it is standard practice to remain true to the formatting of the original text. Rare examples when relative positioning does have linguistic meaning include  .

A further complication is that different intended applications call for different levels of information to be present in hieroglyphic encodings. Examples of applications include:

- Studies in palaeography and epigraphy.
- Study of the translation of a particular text.
- Study of grammar.
- Lexicography.

For palaeography and epigraphy, one would wish to preserve as much as possible of the physical appearance of signs as well as their relative positioning.

For translations, a normalised hieroglyphic rendering is usually sufficient. Where there is doubt about its accuracy, one may wish to compare it to a facsimile of the original manuscript. It is easy to find the relevant fragment of the facsimile on the basis of the normalised rendering provided the latter preserves an appropriate amount of the formatting of the original.

For the study of grammar, much of the appearance of hieroglyphs and their relative positioning are of little relevance. Nevertheless one wishes examples in a grammar book to conform generally to conventions of hieroglyphic composition in order to give an accurate impression of the written language.

In lexicography, the attempt is usually made to abstract away from the formatting of particular instances of words. Ancient Egyptian lacks a notion of orthography in the sense of having a single correct writing, and one word may be written with different sequences of hieroglyphs, even within a single text. Consequently, a lemma in a lexicon may consist of an idealised hieroglyphic writing, possibly without any formatting information at all.

The above four example applications illustrate different sets of requirements one may want to impose on an encoding scheme for hieroglyphic text, some with an emphasis on faithfulness to one particular manuscript, others with an emphasis on uniformity across manuscripts.

Other aspects of this discussion include the versatility of encoding schemes and, related to this, the lifespan of encodings. For example, a representation of an hieroglyphic text that is close to a facsimile, with precisely specified scalings and positionings of signs and custom drawings of non-standard glyphs is not very suitable for applications of automatic processing, such as compilation of word lists, automatic transliteration, etc. Such 'pseudo-facsimile' representations also tend to heavily rely on one particular choice of font, and often on one particular software tool offering certain functionality to indicate relative positioning of signs. This severely limits the lifespan of the encodings, as tools and fonts are typically replaced by others after a relatively short time.

However, it is not self-evident that the lifespan of pseudo-facsimile encodings is an issue in practice. In a typical scenario, one could compile a faithful encoding of a manuscript, then convert this to a general-purpose graphical format, such as JPEG or PDF. This can be included in a publication of the manuscript. Thereafter one may safely discard the encoding as it has few other uses.

In this article we will consider encoding from an entirely different perspective, namely that of creating and maintaining a corpus of hieroglyphic texts that has a reasonable life expectancy and can be used for various applications. These applications are numerous: not only the publication of the texts themselves, in electronic format or on paper, but also the reuse of the material in learning and teaching, extraction of sentences for the use in grammar books, extraction of words for use in lexicography, etc.

Some requirements for such an encoding scheme with both longevity and versatility are:

- stability,
- high expressive power,
- font-independence,
- simplicity,
- precision of meaning, and
- flexible formatting.

The need for an encoding scheme that is stable is obvious. In a large corpus that is under development, it would be impractical if frequent modifications to the corpus were required as a result of changes to the encoding scheme. Connected to this is the need to make the encoding scheme powerful enough to deal with most if not all texts that one may reasonably expect to encounter.

Due to the open-ended nature of hieroglyphs, there is no hope of compiling a 'complete' sign list. However, one would expect the expressive power of the encoding scheme to at least cover most if not all kinds of relative positioning that one finds in practice.

A hieroglyphic font is generally a stylised idealisation of the signs that can be found in good monumental inscriptions. Due to the large diversity of styles across periods and regions, it is unlikely that one font will ever satisfy all scholars. Furthermore, a detailed font with fine lines may be more suitable for printing on paper whereas a less detailed font with thick lines may lend itself better to use on computer monitors. In order to use an encoding in a wide range of applications, it should therefore be independent of a particular font.

Data tends to outlive the software by which it is created. Often this is because programming languages can become obsolete very quickly. It is therefore necessary to use simple data formats for which new processing software can be developed easily. The correctness of this software can be guaranteed if the meaning of constructions in the data formats is precisely defined.

Lastly, some applications, such as alignment with transliterations, require provisions for automatically inserting whitespace within hieroglyphic encodings. However, the encoding scheme

itself should be free of physical linebreaks and pagebreaks, leaving it to each application to determine appropriate places for these.

As illustrated by examples in the following sections, the issue of the sign list cannot be seen as independent from the issue of formatting, at least in many existing encoding schemes. In many cases, inadequacies in operators for relative positioning have led to addition of spurious variant glyphs or combinations of signs. In addition, applications ranging from pseudo-facsimile reproduction to lexicographical analysis are also relevant to these encoding questions: not only which signs (characters or glyphs) should be included, but also what kinds of relative positioning need to be available.

2. WHY THE MANUEL DE CODAGE IS INADEQUATE

It is difficult to talk about a single Manuel de Codage (MdC) encoding of hieroglyphic. This is because the last published version was from 1988 (Buurman *et al.* 1988), henceforth referred to as MdC88. Since then many features have been added to hieroglyphic editors but without proper documentation. Some of these editors were developed by the CCER. One phrase on page 15 of Buurman *et al.* (1988) is particularly revealing:

> [...] the Glyph programme [sic], linked to this enterprise from the beginning, has been improved, which had to be included in the Manual

This suggests that the MdC was not intended as a standard in itself, but rather as a manual for a particular tool. In addition, there are by now many competing hieroglyphic editors, each adding its own features and interpreting various imprecisely documented features from MdC88 in different ways.

Rather than directly criticising the MdC or any of its dialects, it is perhaps more appropriate to criticise the tradition of hieroglyphic encoding starting with Buurman *et al.* (1988). The most serious defects within this tradition are:

- The encoding schemes are specific to particular versions of particular tools.
- The emphasis is on creating pseudo-facsimiles. Long-term storage of hieroglyphic encodings for diverse usage and for reuse has low priority.
- Connected to this, the font used is the one that came with the tool. Exchanging one font with another is not guaranteed to give a satisfactory appearance.

A case in point is the operator `&`. It is not part of MdC88, but it has been part of implementations of Glyph for a long time. It occurs in the expression `G14&X1` in an unfinished, updated Manuel de Codage by Hans van den Berg (1997). The operator can be used to separate two or more occurrences of hieroglyphs. Its meaning is undefined except for a finite set of sequences of hieroglyphs specific to the hieroglyphic editor. Where this meaning is defined, it is a particular relative positioning and/or scaling of the individual hieroglyphs. It is typically used where the two operators `:` for vertical and `*` for horizontal combination do not suffice.

The problem is that the number of combinations of glyphs for which the `&` is needed is potentially unbounded. To put it in another way, if we define an expression with `&` for every occurrence of a hieroglyphic group that cannot be described as purely horizontal or purely vertical arrangement of subgroups, then encoding any new text will require defining new expressions. This makes the encoding scheme unstable to the extreme.

group	EGPZ	RES
	G39&N5	insert[te](G39,N5)
	G39&N29	insert[te](G39,N29)
	G39&X1	insert[te](G39,X1)
	G36&X1	insert[te](G36,X1)
	I10&D58	insert[b](I10,D58)

Table 1. Groups that are not formed by purely horizontal or vertical arrangements, their expressions in the EGPZ, and their expressions in RES (see §3)

Tab. 1 shows a few examples of expressions with `&` out of the no less than 400 such expressions included in the EGPZ (Saqqara Technology 2008). This is of course nowhere near an exhaustive list of combinations of glyphs for which the operators `:` and `*` do not suffice. The problem is the lack of power of the latter two operators, in combination with a possible misconception that horizontal and vertical relative positioning would be the norm in hieroglyphic writing, and other types of relative positioning would be the exception. Even a cursory glance at a few original hieroglyphic inscriptions will immediately refute this assumption, as the so called 'special' groups are very common.

Table 2. The risk of hard coding of scaling factors and absolute positions. What may look satisfactory with one font (left) may be entirely unsatisfactory with a different font (right)

Some dialects of the MdC have tried to solve this problem with hard coding of a scaling factor and an absolute position for each occurrence of a hieroglyph in a 'special' group. The problem with this is that the life expectancy of such an encoding does not extend beyond the lifespan of the font with which the choice of scaling factors and positions were determined. This is illustrated in Tab. 2, assuming two different fonts in which the sun-symbol has different sizes.

The Manuel de Codage has more shortcomings, such as the lack of standardisation and the cumbersome syntax, which make it difficult to develop parsers and renderers. It is also problematic that the Manuel de Codage was designed as a holistic file format, to be used for document preparation, including operators for hard linebreaks and pagebreaks. Had the MdC been restricted to just hieroglyphic encoding to be used within arbitrary file formats, it would have inspired more flexible usage, for example for automatic analysis and lexicography.

Some of these shortcomings can be fixed to a certain extent. For example, one could imagine that the Egyptological community as a whole would at some point agree on a common standardised dialect of the Manuel de Codage. However, the traditional emphasis on pseudo-facsimiles and the assumption that encodings are discarded after publication of a text have had too great an influence on the development of common MdC dialects. A substantial paradigm shift is needed to arrive at an encoding scheme that offers any hope that text encodings might survive a change of font or a change of hieroglyphic rendering tool.

3. PROPOSED SOLUTION

The Revised Encoding Scheme (RES) was introduced in Nederhof (2002) and criticism on it was addressed in Nederhof (2008). The development took place in three stages.

First, we investigated large amounts of hieroglyphic texts, as well as modern (hand-drawn) transcriptions of hieroglyphic and hieratic texts. The purpose of the latter was to find out which aspects of formatting of hieroglyphic texts Egyptologists typically want to preserve. We have deliberately ignored typeset hieroglyphic texts, as those are commonly fettered by technological limitations of the formatting and printing tools that were used.

In the second step we designed a small set of operators to express relative positioning of hieroglyphs, such that in principle all of the 'special' groups we found in real texts can be expressed using a combination of those operators. This has been done without too much concern for the technical difficulty of the implementation of the operators.

The technical realisation came in a third step. Whereas the implementation of the most innovative operators can be difficult, it should be pointed out that this task needs to be done only once, and is outweighed by the ease with which texts can be encoded and the ensuing longer lifespan of encodings, independent of any font.

The font-independence comes from the design decision that the meaning of operators should match observable arrangements of signs. For example, one use of the `insert` operator corresponds to the intuitive arrangement that can be described as 'one sign is to be placed in the free upper-right corner next to another, and scaled appropriately'. Encoding can thereby be done by visual inspection rather than by dragging images by the mouse. The consequence is that the unfortunate situation in Table 2 is avoided.

group	Unicode	RES
	D57	stack[on](D56, T30)
	M3a	stack(G17,M3)
	N18b	insert[scp=0.0](N18,O34)
	N25a	N25[yscale=0.8]
	T3a	T3 :[scp=0.0,fit,fix] N26
	T33a	stack[under](T33, S29)

Table 3. Groups that have been given their own code points in Unicode 5.2, but that can be described equally well by RES expressions

Tab. 1 already presented examples of the use of the `insert` operator. Tab. 3 presents further examples of groups of signs that can be expressed in terms of combinations of more elementary signs using RES operators. These groups have in fact been given explicit code points in Unicode 5.2. By our reckoning, there are 105 such groups out of the 1071 hieroglyphic code points in Unicode 5.2 (Nederhof 2011). This strongly suggests that future extensions of the sign list can remain much more modest and manageable if an encoding scheme such as RES is adopted in place of MdC. It should further be pointed out that overly large sign lists with large portions of extraneous signs and sign combinations, such as the EGPZ mentioned in §2, place an unreasonable and unnecessary burden on font developers.

There are provisions in RES for fine-tuning aspects of the formatting, such as an indication that the distance between two signs should be, say, half or double what it would normally be. This can be used for pseudo-facsimile representations, which may be ill-advised for all but a few applications. One may deliberately want to avoid this type of fine-tuning for most applications. If such fine-tuning is used, it will under normal circumstances not be invalidated by a change of font in the sense that a 'wrong' rendering as in Tab. 2 would be produced.[1]

Lastly, it should be pointed out that great advancements towards more powerful hieroglyphic encoding schemes were already made in PLOTTEXT (Stief 1985). In that system there are, for example, operators for placing a sign in a free corner next to another sign, comparable to our `insert` operator.

4. DISCUSSION

The creation of large electronic corpora of hieroglyphic texts is only cost-effective if the validity of the encodings can be preserved over a long period. In the tradition of the Manuel de Codage, the validity of an encoding is specific to a certain choice of software package and font, which precludes longevity of the electronic resources. Consequently, if there is to be any hope of developing comprehensive corpora, the Egyptological community should abandon the Manuel de Codage encoding of hieroglyphic text. One viable alternative in the form of RES is readily available.

There are currently no well-defined criteria by which one can decide which new hieroglyphs should be added to the Unicode set. Developing such criteria is all the more difficult as the character/glyph dichotomy seems to be far apart from the way that hieroglyphic texts are commonly transcribed, for most relevant applications. It is also possible that systematic investigations of shapes and meanings of signs, such as those by Meeks (2004), will one day brings us closer to an answer. What does seem clear is that a well-designed encoding scheme will avoid the need for extraneous signs, added just to compensate for the inadequacies of the relative positioning operators.

Abstract

In this paper, we discuss the encoding of hieroglyphic text and argue that the set of requirements for an encoding scheme depend on the intended application. Our main claim is that if this application is the development of text corpora with long lifespans and diversity of use, then encoding schemes within the tradition of the Manuel de Codage are unsuitable.

BIBLIOGRAPHY

van den Berg, Hans. 1997. Manuel de Codage: A standard system for the computer-encoding of Egyptian transliteration and hieroglyphic texts, in: http://www.catchpenny.org/codage/ (accessed 2011-09-30).

Buurman, Jan, Nicolas Grimal, Michael Hainsworth, Jochem Hallof & Dirk van der Plas. 1988. *Manuel de codage des textes hiéroglyphiques en vue de leur saisie sur ordinateur*, Paris, Institut de France.

Daniels, Peter T. & William Bright (eds.). 1996. *The World's Writing Systems*, New York, Oxford University Press.

Everson, Michael. 1997. Proposal to encode basic Egyptian hieroglyphs in Plane 1, in : ftp://std.dkuug.dk/jtc1/sc2/WG2/docs/n1637/n1637.htm (accessed 2011-09-30).

———. 1999a. Encoding Egyptian hieroglyphs in Plane 1 of the UCS, in: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n1944.pdf (accessed 2011-09-30).

———. 1999b. Response to comments on the question of encoding Egyptian hieroglyphs in the UCS (N2096), in: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2132.htm (accessed 2011-09-30).

Gardiner, Alan H. 1928. *Catalogue of the Egyptian Hieroglyphic Printing Type, From Matrices Owned and Controlled by Dr. Alan H. Gardiner*, Oxford, Oxford University Press.

1 More information on RES can be found at: http://www.cs.st-andrews.ac.uk/~mjn/egyptian/res/

———. 1929. Additions to the new hieroglyphic fount (1928), in: *The Journal of Egyptian Archaeology* 15, p. 95.

———. 1931. Additions to the new hieroglyphic fount (1931), in: *The Journal of Egyptian Archaeology* 17, p. 245-247.

———. 1953. *Supplement to the Catalogue of the Egyptian Hieroglyphic Printing Type, Showing Acquisitions to December 1953*, Oxford, Oxford University Press.

———. [3]1957. *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*, Oxford, Griffith Institute.

Grimal, Nicolas, Jochen Hallof & Dirk van der Plas. 1993. *Hieroglyphica*, Utrecht-Paris, Publications Interuniversitaires de Recherches Égyptologiques Informatisées.

Meeks, Dimitri. 2004. *Les architraves du temple d'Esna. Paléographie*, Cairo, Institut français d'archéologie orientale (= Paléographie hiéroglyphique 1).

Nederhof, Mark-Jan. 2002. A revised encoding scheme for hieroglyphic, in: *Proceedings of the 14th Table Ronde Informatique et Égyptologie*. On CD-ROM.

———. 2008. Automatic alignment of hieroglyphs and transliteration, in: Nigel Strudwick (ed.), *Information Technology and Egyptology in 2008, Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists*, Gorgias Press, p. 71-92.

———. 2011. The 1071 hieroglyphs from Unicode 5.2, in: http://www.cs.st-andrews.ac.uk/~mjn/egyptian/unicode/ (accessed 2011-09-30).

Saqqara Technology. 2008. EGPZ 1.0 specifications, in : http://www.egpz.com/resources/egpz.htm (accessed 2011-09-30).

Schenkel, Wolfgang. 1999. Comments on the question of encoding Egyptian hieroglyphs in the UCS, in: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2096.pdf (accessed 2011-09-30).

Stief, Norbert. 1985. Hieroglyphen, Koptisch, Umschrift, u.a. – Ein Textausgabesystem –, in: *Göttinger Miszellen* 86, p. 37-44.

Unicode Consortium. 2010. Egyptian hieroglyphs, in http://www.unicode.org/charts/PDF/U13000.pdf (accessed 2011-09-30).