Next steps on Book Pahlavi
Roozbeh Pournader (WhatsApp) and Liang Hai
April 28, 2020

Background
Book Pahlavi may be the best-known complex script not yet encoded in Unicode. It has widespread usage among scholars of Iranian languages, but there are still issues to be resolved before the script can be encoded in Unicode. This document is based on a review of most recent documents in the Book Pahlavi Topical Document list at https://unicode.org/L2/topical/bookpahlavi/, as well as further email communications with Anshuman Pandey.

This document establishes some questions that need to be answered before Unicode can encode the script. The main open question, “What is the right model to encode the script?” remains unanswered. The authors confess they don’t have an answer yet, but believe the information requested here would help arrive at the best model or make large advances towards it.

We consider Pandey 2018 (as opposed to older proposals such as Pournader 2013 and Meyers 2014) to be the baseline further proposals should be based on, as it’s the most comprehensive proposal yet submitted. But we make reference to the older proposals to point open issues.

Technical questions that need answers from experts
1. Meyers 2014, p. 11, mentions the following two specific forms that don’t appear to be described in Pandey 2018:

   \[
   \begin{align*}
   \text{\textcircled{~}} & \quad - & \quad \text{right} & \quad \text{yh}/1 \\
   \text{\textcircled{c}/\textcircled{c}} & \quad - & \quad \text{non-joining} & \quad \text{c}/\text{j}
   \end{align*}
   \]

   Do such forms actually exist in Book Pahlavi texts? If yes, how should they be analyzed? For example, should Meyers’s “yh/1” be analyzed as a sequence of \textit{gimel-daleth-yod} and another character? (If yes, which character?) Should their non-looped non-joining “c/j” (note a looped right-joining c/j/p already exists in the proposal) be considered just a variant of \textit{sadhe} or does it have important distinctions from the looped forms of \textit{sadhe}?

   Note that Meyers 2014, p. 50 includes images of a typeset Book Pahlavi text that appear to show the first glyph, although the examples are not right-joining as
Meyers claims. Are these properly typeset, or are they artifacts of the type?

2. Meyers 2014, p. 15 considers what Pandey 2018 and Pournader 2013 propose as the letter he a digraph:

The specific shape presented includes a protruding part, circled above. Is that a feature of Book Pahlavi or just an artifact of the typesetting technology or a typo? Note that Pandey 2018 also includes such a protrusion on page 21:

the sequence or mem + nun

A similar examples happens on Pandey 2018, page 23:
If such variation actually exists in Pahlavi texts, is it systematic? Does the presence of the protrusion hint toward a specific reading of the text, such as $mn$ as opposed to $h/E$?

3. Meyers 2014, p. 18 mentions three diacritics that are not proposed in Pournader 2013 (which is based on Nyberg 1964). These are caron below, dot above, and three dots below. They claim it appears in Katāyūn Mazdāpur's Dāstān-e Garšāsp, Tahmūres o Jamšīd, Gelšāh o Matnāh-ye dīgar.

   - Circumflex (蹙) [15, p.123]
   - Caron below (蹙) [15, p.123]
   - Dot above (蹙) [15, p.125]
   - Dot below (蹙) [15, p.122]
   - Two-dots above (蹙) [15, p.124]
   - Two-dots below (蹙) [15, p.124]
   - Three-dots above (蹙) [15, p.124]
   - Three-dots below (蹙) [15, p.124]

Pandey 2018 includes all these diacritics, but reverses the caron below, calling it a "hat below":

   - Caron below (蹙)

Do such diacritics indeed exist? If yes, which letters are they used with? What is the phonetic value of the dotted letters?

4. Meyers 2014, p. 39, Figure 4.3, contains some words or letters in white frames. What is their reading? What do the dot diacritics indicate?
5. Meyers 2014, pp. 46–48, contains some four-dotted and multi-dotted punctuations. Are these common or rare (or hapaxes)? Does the difference with the common three-dot punctuations signify something?
6. Meyers 2014, p. 47, includes some text at its top right:

Is this Book Pahlavi or another script like Avestan or Arabic? If it is Book Pahlavi, what’s the reading?

7. Meyers 2014 p. 53 shows these two samples:

Are they the same spelling? Could they have different readings? What does the extra tooth in the green word signify? Is this a feature of manuscripts, or something that only appears in typeset texts?

Other issues

8. Provide a complete list of dotted letters: which letters combine with which diacritics and what is their phonetical value? Nyberg 1964 provides the following list:

- \( \text{gimel-daleth-yodh} + \text{two dots above} = \text{g} \)
- \( \text{gimel-daleth-yodh} + \text{hat above} = \text{d} \)
- \( \text{gimel-daleth-yodh} + \text{two dots below} = \text{y} \)
- \( \text{gimel-daleth-yodh} + \text{dot below} = \text{j} \)
- \( \text{shin} + \text{three dots above} = \$ \)

What other combinations are used in Book Pahlavi?

9. Nyberg 1964, p. 135, mentions a hook used under \textit{mem-qoph}. How is that analyzed? Does it need a separate encoding as a combining mark or an alternate form of \textit{mem-qoph}?

10. What are the basic graphemes of numbers and which of them should be unified with letters? What are the joining properties of numbers? Are there numbers which join to the previous or next character sometimes and don’t join some other times? (See Nyberg 1964 pp. 173–174 for “Figures” and “Ordinals” as well as its page 131 for the
suffix form of the number one. See also Skjærvø 2008 pp. 97–99 and Meyers 2014 Section 2.7.)

Recommendations by authors

- Book Pahlavi should be encoded as a cursive script (using rules similar to Arabic shaping) as proposed by Pandey 2018 and Pournader 2013. Otherwise, unreliable hacking would be needed for implementing ideal cursive-joining fonts. Contextual variations deemed unnecessary by Meyers 2014 (which are actually desirable in most fonts) would not be reliably rendered if it is encoded as non-cursive, as text engines would assume the script is not complex, resulting in subpar renderings.

- If there are indeed two forms of upside-down Ahriman as mentioned by Meyers 2014 Section 2.4.1, two Atomic characters should be proposed for them. Typographic inversion is hard to achieve in modern text processing environments and encoding these as special characters helps the user community avoid complex tricks to typeset common texts.

- Meyers 2014 proposes encoding a smoother form and a squarish form of h/1 as two different characters. They should not be disunified. The variation is predictable in contexts (e.g. becoming squarish when having letters inside the belly), see Meyers 2014 p. 38 Figure 4.2.

Alternatively, it may be the case that numbers have a different curvature than letters (if the cyan box is showing numbers), and in such a case, numbers should be disunified from letters.

- Meyers 2014 p. 17 Section 2.5 talks about occasional letter separation and recommends the usage of U+304F COMBINING GRAPHEME JOINER (CGJ) for such cases. This is inconsistent with other usages of the CGJ. Instead ZWNJ should be used to break joining, or one of the various thin spaces could be used if more space is needed.

- Pandey 2018 introduces the concept of “fixed-form” letters, to work around cases where “normal joining behavior is suspended”. While we agree that a good case is made for such characters, we think the exact model proposed by Pandey will create confusion and ambiguity. We think that instead of one “normal” letter and one “fixed-form” letter, we should consider two “fixed-form” letters, one that always forms a belly and another that never does. In this way, users of the script and font
designers wouldn't need to learn the complex rules on when a belly is formed, and text is more predictable while being typed. (Note that this may result in a reduction in the number of letters in total, since some of these fixed forms may need to be unified with other characters.)

Ambiguity and encoding

It is well-known that Book Pahlavi texts are quite ambiguous. This is indeed the source of the radical encoding model proposed by Meyers 2014, which tries to resolve those ambiguities and reduce Book Pahlavi text to very basic elements, which in some cases lose their relation to letters.

As mentioned earlier, the authors do not have a complete model in mind, and the ambiguities of Book Pahlavi concerns them too. The right model may live somewhere between the model of Pournader/Pandey on one side and Meyers on the other. What follows, is some of our thoughts about what may be the best model.

First, there are some very basic cases of ambiguity in Book Pahlavi, which are very frequent. These two come to mind:

- The confusable belly part in shin versus the bellied form of aleph; and
- samekh vs two consecutive gimel-daleth-yodhs.

Generally, we think in order to reduce ambiguity in the encoding, phonemic spelling should be deprioritized, and the graphetic display of the text should become the primary source of encoding decisions. For example, if a phonemic misspelling of some word would result in the right display, it may be preferable to represent the word using that misspelling, as opposed to a phonemically correct spelling that would result in a different display. This line of thinking leads to a need for investigating whether we can reduce the number of letters encoded while keeping every known word representable.

Alternatively, in a more phonetic model like the one proposed by Pandey 2018, we recommend using the most simple pieced-together representation according to how a word appears. For example, consider the case Meyers 2014 p. 32 brings up, where the words gy’h and sy’d are spelled the same way visually:

The authors have confirmed this double reading by checking with MacKenzie 1986 (p. 167), which also gives their spellings as gy’h (p. 36) and sy’d (p. 78):

\[\text{giyāh, syā} \]

Assuming Meyers’s analysis of the elements of the word is correct, perhaps the word should be encoded as it appears on the page, say gy’h, regardless of the phonetic/semantic interpretation, using the spelling that does not require the font know about an alternate curl-less form of y for the spelling sy’d.
An example of this more visual model would be handling the confusion between the final forms of *pe* and *sadhe*. The right part of *pe* frequently (but not always) merges into a preceding stroke, making it indistinguishable from the final form of *sadhe*. In such cases, the words using such forms of *pe* should be encoded using *sadhe*.

There are some more complex situations, such as Meyers 2014’s claim in its page 16 that a curved belly in some *<aleph-heth, taw>* sequences could carry semantic information, as opposed to say an *<aleph-heth, aleph-heth>* sequence. This is in contrast with Pandey 2018, which considers the situations to be the same. This needs further investigation.

Finally, we believe Unicode should try its best to avoid sub-letter encoding. We find the mismatch between letter boundaries and character boundaries (Meyers 2014, p. 33) concerning.

**Bibliography**