

Proposal for Encoding Book Pahlavi in the Unicode Standard

Version 1.2

Abe Meyers¹

May 9, 2014

¹abraham DOT meyers AT orientology DOT ca

Foreword

The purpose of this document is two-fold. First, to propose the inclusion of the Book Pahlavi script characters in the Unicode Standard, and second, to outline and address some of the shortcomings of the previous proposed encoding models [7, 20], in particular the more recent of the two [20].

The previous proposal admits that it “does not attempt to solve the multi-layer and complex problems of properly and completely representing text written in Book Pahlavi” [20, p.3]. The current proposal, on the other hand, proceeds to indeed solve those very problems.

My first goal is to be able to uniquely and unambiguously represent Book Pahlavi texts in Unicode, without any loss of vital information in the transcoding process. To elaborate, first, a given *shape* standing for a Pahlavi word, should be able to be encoded only in *one* way to Unicode. Second, starting from the encoded word, only *one* shape must be able to be generated. Third, going through a round trip, i.e. encoding the word and then rendering it, the rendered shape and the original shape should contain the same amount of (relevant) information.

In addition, I will encode the text in such a way that in the process of rendering it, no obligate ligature features and no obligate contextual shape changes, and in general no complex text layout (CTL) capabilities, are needed to get “minimum legible” text as defined by the Unicode Standard. All these goals are met while the total number of the proposed letter-like characters (vs diacritics and punctuation marks) is *equal* to that of the previous proposal. Furthermore, my proposed method will eliminate the need for using multiple fonts, or high-level markup, or multiple glyph features of a single font to adequately represent Book Pahlavi texts.

The differences between the character repertoire in this proposal compared to what is in the previous proposal [20] can be summarized as follows:

- Removal of digraphs as separate characters
- Addition of new basic characters

- Addition of new diacritics
- Addition of new punctuation marks

Contents

1	Introduction	8
2	Book Pahlavi Characters	10
2.1	Basic characters	10
2.1.1	Notes on variant characters	12
2.1.2	Dealing with corrupt forms	13
2.2	Joining behaviour	14
2.3	Digraphs	14
2.4	Ligatures	15
2.4.1	The case for Ahreman	15
2.4.2	Stylistic and aesthetic ligatures and kernings	15
	Extra curvings when attaching to \mathfrak{C}	15
	Extra curvings when attaching to \mathfrak{P}	16
	\mathfrak{C} Ligatures	16
	\mathfrak{P} Ligatures	16
	Vertical kerning and ligature for \mathfrak{A}	17
2.5	Occasional letter separation	17
2.6	Diacritics	18
2.7	Numerals	18
2.8	Kashida	19
2.9	Punctuation	21
2.10	Proposed character mapping in Unicode	21
2.11	Sorting	25
2.11.1	Collating basic characters	25
	Example	27
2.11.2	Collating diacritics	28
2.12	Text standardization and normalization	28

<i>CONTENTS</i>	4
3 Previous Encoding Models Problems	30
3.1 Encoding to Unicode	31
3.2 Decoding and rendering from Unicode	33
4 Sources and Examples	36
4.1 Samples from different sources	36
4.2 Encoding example	61

List of Tables

2.1	Basic characters in Book Pahlavi	11
2.2	Book Pahlavi digraphs	15
2.3	Book Pahlavi numbers	20
2.4	Proposed mapping to Unicode	22
2.5	Character names for the Unicode Standard	23
2.6	The fragment of <code>UnicodeData.txt</code> pertaining to Book Pahlavi	24
2.7	Sorting keys for characters digraphs and other character combinations	26
2.8	Applying collation rules of table 2.7 to a list of words	27

List of Figures

2.1	Sorting order of the basic characters	27
4.1	Examples of use of x_1 and x_2 and the genitive preposition . .	37
4.2	A sample page of <i>Bundahishn</i> (TD2)	38
4.3	A page from a manuscript that has ordered letters	39
4.4	A page from a manuscript of <i>Minug-i Xrad</i>	40
4.5	The first page of a manuscript of <i>Sifat-i Siroozeh</i>	41
4.6	A page from a Manuscript with Persian transcription	42
4.7	A page from a Manuscript depicting assorted numbers from 1 to 100	43
4.8	A page from a Manuscript depicting assorted numbers from 500 to 70,000	44
4.9	A page from a Manuscript depicting assorted numbers from 80,000 to 1,000,000	45
4.10	A page of a Pahlavi glossary from a manuscript	46
4.11	A page from a Manuscript showing the 4-dot punctuation mark	47
4.12	Fragments of a manuscript showing punctuation marks	48
4.13	A fragment from a manuscript showing ordinal 13th	49
4.14	Variants of numerals from 2–17 using different ciphers	50
4.15	Variants of selected numerals from 18–10,000 using different ciphers	51
4.16	A page of Vendidād and its Zand	52
4.17	A page from Book III of <i>Dinkard</i>	53
4.18	A page from Jamasp-Asana’s Pahlavi corpus	54
4.19	A typeset fragment from <i>Zand of Barān Yasn</i>	55
4.20	A typeset fragment from <i>Zand of Srōš Yasn</i>	56
4.21	A passage from <i>Minug-i Xrad</i> handwritten by contemporary scholars	57
4.22	A page of typeset Pahlavi text from <i>Zand of Bahman Yasn</i> .	58

LIST OF FIGURES

7

4.23 A page from the glossary of <i>Zand of Bahman Yasn</i>	59
4.24 A Page from MacKenzie's Pahlavi Dictionary	60

Chapter 1

Introduction

Book Pahlavi is a script that was used for writing books in Middle Persian. The majority of the Zoroastrian religious texts written in Middle Persian are written in this script. Book Pahlavi is considered the cursive counterpart of the Inscription Pahlavi script. Through the passage of time, the cursive off-shoot evolved into a separate script where the shapes for many characters changed, and furthermore, several characters of Inscription Pahlavi were reduced to a single character in the Book Pahlavi script—adding more ambiguity. There is extant material written in Book Pahlavi that belongs to the late Sassanian period, such as letters or seal inscriptions. However the majority of the existing Pahlavi books (in the conventional sense of a physical book) date back from the 9th to the 11th century AD. The extant manuscripts in general do not go back further than the 14th century.

The script, like the other descendants of Aramaic scripts, is written from right to left. The cursive nature of Book Pahlavi frequently results in characters joining one-another, similar to the case of cursive English handwriting. Although, some characters do not join their preceding character, almost all of them allow the terminal stems of the previous character (i.e., the one to their right) to join them—provided that the character to the right *can* join its next character.

A peculiar property of the texts written in all variants of Pahlavi scripts, including Book Pahlavi, is the presence of Aramaic ideograms. These are words that are written in Aramaic (using the Pahlavi script) but when being read, the reader would substitute the word with the Persian equivalent. For example, they would write 𐭮𐭥 ($MN < \text{Aramaic } mn$) which means *from* in Aramaic, but would read it as *az*, which means *from* in Persian.

These Aramaic ideograms, called *Huzvārišn* in Middle Persian, can freely

be combined with Middle Persian affixes and words in Pahlavi texts. For example, the word 𐭠𐭥𐭥𐭥 (*MLKA*) which is the ideogram for *šāh* (= *king*) can combine with the Persian pluralization suffix *ān* (𐭠𐭥) to form 𐭠𐭥𐭥𐭥𐭠𐭥, transliterated as *MLKAn MLKA* and read as *šāhān šāh*, meaning *king of kings*.

The Book Pahlavi script is extremely ambiguous and there are numerous occasions where scholars disagree on how to read a single word. There are several sources of ambiguity. First, a single character can stand for any of multiple phonemes, e.g., 𐭠 can stand for /g/, /d/, or /y/ phonemes. Second, there are multiple digraphs in Book Pahlavi.¹ For example, 𐭠𐭠 can stand for the phoneme /s/, in addition to a pair of phonemes with each of them corresponding to one of /g/, /d/ or /y/, e.g., *gg*, *gd*, and etc.. Third, the short vowels are mostly not present in Pahlavi orthography. Furthermore, the scholars should decide if they are dealing with Middle Persian words or their Huzvārišn. Finally, add to the mix, the historic and pseudo-historic orthography and numerous orthographic inconsistencies across different manuscripts.

The common practice among the Middle Persian scholars, when dealing with a piece of Pahlavi text, is to first transliterate the text into an intermediate form and then transcribe the transliteration to phonemic Pahlavi. That is, to transcribe the transliteration letters to proper phonemes and add the missing vowels and convert the ideograms to their Middle Persian equivalent forms. It is at the step of the transliteration that the scholar should decide if he is dealing with a Middle Persian word or its Semitic ideogram. The Semitic ideograms are transliterated to capital letters. Regular text is transliterated to lowercase characters.

Book Pahlavi scribes were aware of the script's ambiguity and sometimes they would write the phonetic transcription of a word in the extremely phonetic Avestan script for clarity. This practice was called *Pāzand*. There are a handful of Middle Persian texts that are completely written in *Pāzand*. In technical religious texts, one frequently encounters technical and potentially non-familiar terms written in *Pāzand* form, amidst the Book Pahlavi text. Another example of mixing Pahlavi and Avestan scripts is in the annotations of the Avestan text. The passages from the sacred book are written in the Avestan script and the annotations called *Zand* are written in Book Pahlavi. See figure 4.16 for an example.

¹ *sh* is a familiar example of a digraph in English which can represent the phoneme *š* in addition to two consecutive phonemes of say /s/ and /h/

Chapter 2

Book Pahlavi Characters

2.1 Basic characters

The script has 20 canonical basic character forms, which are listed in table 2.1 along with their common transliteration. The transliteration list is by no means exhaustive and is only provided to add context. A couple of the characters have variants (alternate forms) that are listed in the table as well.

Looking at the table 2.1, a few comments need to be made. The character 𐭌 while transliterated as *l*, may represent /l/ or /r/ phonemes. To remedy this ambiguity, in some manuscripts they have introduced 𐭍 to uniquely represent /l/ and remove the possibility of /r/. In some manuscripts instead of the tick a small circle is used, but that is just a stylistic variation of 𐭍.

𐭎, transliterated as *L*, is only used in Huzvārišn, and is used at the end of the Huzvārišn word [31, p.128]. *L* in other parts of the Huzvārišn word and also sometimes at the end of the Huzvārišn word is written as 𐭌, e.g., 𐭌𐭎𐭎 (ZKL) meaning male [2, p.144], although the form 𐭎𐭎 is also used [31, p.128]. If the 𐭎-ending Huzvārišn word itself is combined with a Persian suffix then 𐭎 retains its shape, e.g., 𐭎𐭎𐭎 (AHL-yh) which transcribes to *pasīh* meaning *rear* (see [13, p.235]).

𐭏 and 𐭐 are variant glyphs of the same character. The character is usually used at the beginning of the word or after right-only joining characters. However, it occasionally stands for the Middle Persian suffix -(i)z [13, p.153], in which case it comes at the end of the word, regardless of the character that the word ends with, e.g. in the word 𐭏𐭎 meaning *even to, moreover unto*; see [31, p.248].

As a matter of orthography, the character 𐭑 is not used in isolation or

Character	Variants	Joining side	Common translit.
𐬀	𐬀	right	b/1
𐬁	𐬁	right and left	g/d/y
𐬂	-	right	d/10
𐬃	-	right and left	g/d/y
𐬄	-	right	yh/1
𐬅	-	right	n/w/r
𐬆	-	right and left	z
𐬇	-	right	k
𐬈	-	right	k/K/γ
𐬉	-	right and left	l/r
𐬊	-	right and left	l
𐬋	-	right and left	L
𐬌	-	right and left	m
𐬍	-	right and left	
𐬎	-	right and left	s
𐬏	-	non-joining	c/j
𐬐	-	right	c/j/p
𐬑	-	right	t
𐬒	-	right	x ₁
𐬓	-	right	x ₂

Table 2.1: Basic characters in Book Pahlavi

after characters that do not join from the right.

The character \smile is sometimes used as an alternate form for the digraph $\mu\upsilon \leftarrow \mu + \upsilon$.¹ when used at the end of the word. So in this sense it can be considered a stylistic ligature. The reason that is considered a basic character and *not* a mere ligature is that the character \smile occasionally represents the numerical value 1, used at the end of a numeric compound, as an alternate form of \smile —acting as a cipher. On the other hand, the digraph $\mu\upsilon$ is *never* used as a cipher to represent a numerical value of 1 [31, pp.334–336].

The characters μ and υ have roots in scribal ligatures, but they are mostly used as atomic units in the texts and are treated like other basic characters. They are also used outside the realm of the combinations that they supposedly represent. Most modern scholarly books also assign them their own separate transliteration letters x_1 and x_2 , respectively [2, 19, 22]. The notable exception is MacKenzie which denotes them as \underline{yt} and \underline{ty} [13, p.xiv].

2.1.1 Notes on variant characters

I use the following maxim to decide if a variant (alternate form) of a *single* character (as opposed to a combination of two characters) merits its own Unicode code or if it should be considered a variant glyph at the font level (and not have a separate Unicode character): If both variants of the character are found in the *same* manuscript—and preferably there are multiple manuscripts in each of which both variants of the character are found—then the variants have a *necessary* requirement to have separate Unicode numbers. If the variant helps with disambiguation of the reading, I consider the combination of the two conditions as *sufficient* reasons for the variant character to have its own Unicode code.

On the other hand, if the same character is written slightly differently across different manuscripts but consistent within each manuscript and furthermore it doesn't help with the reading of the text, then I consider the variation as a variation in font.

With the introduction, I posit the following about the variations that are mentioned in table 2.1: The difference between \mathfrak{C} and \mathfrak{C} is a difference in font, therefore \mathfrak{C} does not merit its own character.

¹As a matter of convention in this document, whenever combinations of Pahlavi characters are shown with the plus (+) sign in between them, the whole combination should be read from *right to left*. Combinations of Latin letters should be read from left to right as normally done

∩ is a variant of J and merits its own character. It is used alongside J in many manuscripts and is *solely* used to indicate the genitive preposition /i/ which means *of*, see figure 4.1 for an example.

A similar case goes for 𐭪. It merits its own Unicode character as it has been used alongside 𐭫 in the same manuscript and also helps with disambiguation, as 𐭪 is a cipher representing the value 1. Refer to figures 4.7–4.10 and section 2.7.

2.1.2 Dealing with corrupt forms

Pahlavi manuscripts are notorious for orthographic inconsistencies and letter corruptions. Frequently, similar characters are used instead of one another, and sometimes such usages are consistent throughout a manuscript. For example, 𐭫 may be used instead of 𐭪. In such cases it is transliterated as k. Three approaches can be taken when dealing with the corrupt forms.

1. Use the same Unicode character as the correct character but use a different glyph. This is at the font level, or some markup level.
2. Consider it a character variant in the sense defined in 2.1.1, with a different Unicode character for the corrupt form. In this case the glyph for this character is the same as glyph for the wrong character, but the Unicode character for it is different and distinct.
3. Use the same character and glyph as the wrong character.

Pros and cons can be argued for all sides. The problem with the first approach is that there is information loss when encoding the text into plain text Unicode. The problem with the second approach is that a given piece of text written on paper can be encoded differently based on how the scholar reads or interprets it. The first approach also has this problem. The problem with the third approach is that a given Pahlavi word can have several spellings and that can pose problems or extra work in constructing corpi and processing the text.

Ultimately I have taken the third approach because I believe the problems with the first two are far greater and non-negotiable. Furthermore in Iranology circles most standardized material (to be used by pupils for example) fix these corruption cases anyway.

2.2 Joining behaviour

The joining behaviour of Pahlavi characters are far easier than that of complex scripts such as Arabic, and is much closer to that of cursive variants of say English. Contrary to complex scripts like Arabic, there is no fundamental notion of positional shape change of characters in my proposal. The characters look more or less the same regardless of their position. Of course there are many stylistic ligatures that are frequently (but not always) applied to pairs of characters. These include curving the stems and adding horizontal and vertical kernings. I will touch on some of them in subsequent sections. All fine details of aesthetically joining two characters should be handled by the font, through kerning tables and ligature pair tables, in case of OpenType fonts. Therefore, the information provided under the “Joining side” column of table 2.1 is just a guide for typeface designers to allocate proper whitespace around their glyphs.

The joining behaviour is simple because I have chosen an *irreducible* set of Basic Book Pahlavi characters. If I start with a reducible set, like what was proposed in [20], then there are multiple obligate ligature and context-dependant shape changes that will require a complex joining behaviour like that of Arabic. Unfortunately, when starting with a reducible set even having Arabic joining will not solve all problems. I elaborate on this issue in subsequent sections; in particular see section 3.1.

2.3 Digraphs

Book Pahlavi has five digraphs which I have listed in table 2.2. Note that as in the case of *sh* in English, no new Unicode character is needed to represent these digraphs. Refraining from encoding digraphs separately is inline with Unicode’s Technical Committee (UTC) position that “no new digraphs should be encoded, and that their special support should be handled by having implementations recognize the character sequence and treat it like a digraph” [30].

I remind the reader that throughout this document and in particular in table 2.2, Pahlavi letter combinations (using the sign +) are read from right to left.

Digraph	parts	Common Transl.
𐭪𐭪	𐭪 + 𐭪	h/ ʾ
𐭪𐭫	𐭪 + 𐭫	E
𐭪𐭬	𐭪 + 𐭬	p
𐭪𐭭	𐭪 + 𐭭	š
𐭪𐭮	𐭪 + 𐭮	s

Table 2.2: Book Pahlavi digraphs, representing new phonemes

2.4 Ligatures

2.4.1 The case for Ahreman

The word denoting Ahreman (Zoroastrian evil deity) is written upside down in manuscripts (180° rotation) as 𐭠𐭡𐭢𐭣 or 𐭠𐭡𐭢𐭣, see figure 4.22 and figure 4.24. I believe that the upside-down forms do not merit their own characters, but rather they should be considered a form of typographic emphasis. The first reason is that the rotation is not required for minimal legibility. In fact there are instances—especially in glossaries—that the words are written regularly. For example see figure 4.23. Second, by the virtue that the word is written upside down in both variants of the spellings, it seems that the act of turning *Ahreman* on its head is applied to the concept and is not a “property” of the script.

2.4.2 Stylistic and aesthetic ligatures and kernings

There are a few stylistic and aesthetic ligatures that were frequently (but not always) employed by Pahlavi scribes. In general, these ligatures convey no extra semantic information and as said they are purely stylistic and aesthetic and none of them are obligate. Their application varies from manuscript to manuscript and they are not required for minimum legibility. In this section I review some of them for the sake of completeness and to provide a general guideline for Pahlavi type designers who want to create high-quality fonts. None of the ligatures mentioned in this subsection merit their own Unicode characters.

Extra curvings when attaching to 𐭪

The horizontal terminal stem of characters are sometimes curved and joined to the 𐭪 for aesthetic reasons. For example:

- $\text{𐭪} \leftarrow \text{𐭬} \leftarrow \text{𐭬} + \text{𐭪}$
- $\text{𐭪} \leftarrow \text{𐭬} \leftarrow \text{𐭬} + \text{𐭪}$

Similar behaviour, meaning extra curving of the combined glyph, may happen for attaching 𐭪 and 𐭫 .

Extra curvings when attaching to 𐭮

Occasionally, the terminal stem of the previous character to 𐭮 is curved to attach nicely to 𐭮 . This is especially true for 𐭪 and 𐭫 . See figure 4.16 (boxed in green) as an example. See figure 4.6 for a case in which these ligatures are not used. These ligature do not offer any extra semantic information. In some books, for cases like 𐭮𐭪𐭮 the extra curving of stem applies to both meaning that I have a ligature made of three characters. However, in a piece of Pahlavi text that I normally have such tri-character ligatures, if the ligature only applies to the combination of the second 𐭪 and 𐭮 , i.e. 𐭮𐭪 , and the stem for the first 𐭪 is not curved, then this ligature conveys extra semantic information. The extra information is that the combination of 𐭮𐭪 in the fragment 𐭮𐭪𐭮 is *not* a digraph and should *not* be transliterated as *s*. In such cases, when encoding to Unicode, the character U+200C, ZERO-WIDTH NON-JOINER (ZWNJ) should be inserted after the first 𐭪 to prevent ligature formation and convey the extra semantic information. For purposes of collation and other computerized text processing, it is recommended that the character U+034F COMBINING GRAPHEME JOINER (CGJ) be inserted before or after ZWNJ. See section 2.5 for a similar process.

𐭪 Ligatures

Another group of stylistic aesthetic ligatures are the ones that pertain to the way 𐭪 attaches to its preceding character. Frequently the glyph 𐭪 is used.²

- $\text{𐭪} \leftarrow \text{𐭪} \leftarrow \text{𐭪} + \text{𐭪}$

When this ligature is applied, in many cases anything preceding 𐭪 is raised so that the horizontal final stem of 𐭪 ends up on the baseline.

𐭫 Ligatures

Sometimes, the tail of the character 𐭫 is extended and curved towards the right before going further down, when it is attaching to 𐭫 . For example, the

²See [14, p12:4] and many other occurrences in which this aesthetic ligature is *not* used.

word $\mu\mu\mu\mu$ is written like $\mu\mu\mu\mu$ in [6]. As it can be seen, the first two instances of μ are the conventional form and the last instance has the extra curvy tail. Of course, another scribe may choose not to apply the stylistic ligature. For example the word $\mu\mu\mu\mu$ is written like $\mu\mu\mu\mu$ in [9].

Vertical kerning and ligature for \mathfrak{g}

Sometimes when \mathfrak{J} or \mathfrak{S} are followed by \mathfrak{g} , the characters are vertically kerned so that the final stem of \mathfrak{g} and that of the preceding character join smoothly. For example: $\mathfrak{gS} \leftarrow \mathfrak{gS} \leftarrow \mathfrak{g} + \mathfrak{S}$. See figure 4.16 for examples (boxed in brown).

2.5 Occasional letter separation

In some high-quality Pahlavi books printed in India in the late 19th century and early 20th century, they sometimes add tiny hair separation between two characters.³ Usually, this extra cosmetic space has no semantic value. For example the word $\mu\mu\mu\mu$ is written like $\mu\mu\mu\mu$ as written in [23]: $\mu\mu\mu\mu$.

Note the space between \mathfrak{J} and \mathfrak{D} .

However, in some cases when the scribe (typesetter) is aware of the pronunciations of the word, the existence of this slight separation can add semantic information and remove ambiguity on how to read the word. For example, the scribe can write the word $\mu\mu\mu$ as $\mu\mu\mu$ to help the reader read it (or in the case of scholar transliterate it) as *dyk'* and not *sg'*. In other words, by adding the narrow space, the scribe has emphasized that the first two \mathfrak{J} 's need to be interpreted as two separate letters and not a digraph and hence not representing the *s* phoneme. On a semantic level, this information can be encoded (for the most part) by inserting the character U+304F COMBINING GRAPHEME JOINER (CGJ) where the tiny separation occurs. However, since CGJ is usually ignored by the rendering engine [5], it might be desirable to visually accentuate such separation. This is usually the job of higher-level typesetting applications (say \TeX). Occasionally the behaviour may be desirable in plain text. I recommend the use (misuse?) of non-breaking thin space (U+202F) in such cases. If the goal is solely to prevent ligature formation then of course ZWNJ (U+200C) can also be used to prevent ligature formation. This is especially true when dealing with cases like $\mu\mu$.

³This could be explained as inadequate typesetting sophistication but as I will see may add semantic value.

2.6 Diacritics





There are eight diacritics used in Pahlavi manuscripts. Not all diacritics are used in all manuscripts and their use is somewhat haphazard. The diacritics are usually used to specify the phoneme that the bare letter is supposed to represent (in the context) from the set of all potential phonemes that the letter can represent. For example 𐬀 can represent *g*, *d* or *y*, but 𐬀̂ stands for *g*, and 𐬀̇ stands for *d* and 𐬀̈ stands for *y*. Below I list Pahlavi diacritics.

- Circumflex (𐬀̂) [15, p.123]
- Caron below (𐬀̃) [15, p.123]
- Dot above (𐬀̇) [15, p.125]
- Dot below (𐬀̈) [15, p.122]
- Two-dots above (𐬀̈̈) [15, p.124]
- Two-dots below (𐬀̈̈̈) [15, p.124]
- Three-dots above (𐬀̈̈̈̈) [15, p.124]
- Three-dots below (𐬀̈̈̈̈̈) [15, p.124]





Two-dot below, Two-dot above, and Circumflex are the most common diacritics used in Pahlavi manuscripts. Other diacritics are used less frequently. The manuscript MU-16 is one of the richest manuscripts when it comes to the use of diacritics. Therefore, when listing the diacritics used in Book Pahlavi, I have included references to [15] which is the annotated scholarly transliteration, transcription and translation of the manuscript MU-16. I have marked some diacritics as examples in figure 4.3 which is taken from another manuscript.







2.7 Numerals







Numbers are represented using a subset of the basic characters introduced above or their alternate forms; sometimes the circumflex diacritic is used as well. As a whole, the practice of representing numbers is similar to that of the Roman numerals. The ciphers used for denoting numerals are 𐬀, 𐬀̂, 𐬀̇, 𐬀̈, 𐬀̈̈, 𐬀̈̈̈, 𐬀̈̈̈̈, and 𐬀̈̈̈̈̈. For example, the number 3 when written in ciphers




(instead of spelling out the word or employing Huzvārišn) is written as  constructed as  +  + .

There are some inconsistencies among sources on how to represent numbers. These in turn stem from the inconsistencies between different manuscripts. West and Haug [31] do a good job of listing the variant forms of numbers, see figures 4.14 and 4.15. I have tabulated some common variant forms in table 2.7.






A common source of inconsistency is representing numbers from 12–19, where sometimes  and sometimes  is used to denote the *eleven* part of the number. For example the number 12 is either written as  or as .

In the case of numbers 10–19, the hats are frequently dropped, e.g., in [14]. Use of  instead of  to represent number 10 should be considered corrupt usage. Some modern sources do not recognize  and use  [2, 13]. On the other hand, some modern scholarly books that have typeset Pahlavi passages do recognize this character. For example, see [17] and figure 4.19. Looking at Iranian and Indian manuscripts and high-quality books written in India, it is clear that  is a different character from . See figures 4.7, 4.8, 4.9, 4.13, and 4.17 as examples.

 is a variant of  which is used to denote the singularity suffix or number one and transliterated as \bar{y} . Usually, there is a significant overlap of shapes between  and , and for all intents and purposes they are the same character. Most sources consider them the same character too [2, 13, 19]. However, in some manuscripts (for example see [3] and figure 4.2 and [26] and figures 4.7 and 4.10) the difference is pretty clear most of the time. , which is used for numbers is more round in shape and  is more angular. Among modern sources [22] recognizes the difference. Of course, having this distinction helps with disambiguation of reading the text too, therefore I have the necessary and sufficient conditions to denote a unique Unicode code to the alternate character; see section 2.1.1.

Finally, as alluded in 2.1, the character  is sometimes used instead of  as the cipher to denote the value of 1. For example the number 5 may be represented as  [31, p.335].

2.8 Kashida

The character U+240 known as Kashida or Arabic tatweel can in particular be used after characters that end with horizontal stems that lie on the base line. These are , , , ,  and arguably \mathcal{S} (depending on the font).

Numeric Value	Pahlavi representation
1	𐬀
2	𐬁
3	𐬂 / 𐬃
4	𐬄
5	𐬅
6	𐬆
7	𐬇
8	𐬈
9	𐬉
10	𐬊
11	𐬋
12	𐬌
13	𐬍
14	𐬎
15	𐬏
20	𐬐
21	𐬑
30	𐬒
31	𐬓
40	𐬔
41	𐬕
50	𐬖 / 𐬗
60	𐬘
70	𐬙 / 𐬚
80	𐬛
90	𐬜
1 × 100 and 100	𐬝 / 𐬞
200 (2 × 100)	𐬟 / 𐬠
1000	𐬡 / 𐬢
2000	𐬣

Table 2.3: One rendition of Pahlavi numbers. There are many variants.

Font designers may introduce other ligatures for changing the behaviour of other characters. *Kashida* has no semantic value except in few cases. In particular it can be used after the combination 𐬀 to explicitly denote the suffix *-ih*; see [31, p.1]. Second, it can be used to shift the position of diacritics in 𐬀 and 𐬀 to the left.⁴

2.9 Punctuation

The punctuation characters in Pahlavi are the same as those in Avestan. The Avestan punctuation characters cover the range 10B39–10B3F in the latest Unicode Standard (6.3) [5]. However, there are punctuation marks used in manuscripts that are not mapped to any Unicode characters. A full investigation of non-encoded punctuation marks in Avestan and Pahlavi scripts is the subject of another proposal. In the current proposal I mention two extra punctuation marks that are not mapped to any Unicode characters. See figures 4.10 and 4.11. Figure 4.12 depicts a few more punctuation marks that have not been encoded in Unicode.

2.10 Proposed character mapping in Unicode

Table 2.4 shows my proposed mapping between Unicode and Book Pahlavi characters. In table 2.5 character names for the sake of the standard are proposed. I followed [20] in determining the starting point of the first character.

I were not sure where to put the new punctuation characters as there is not much room left around the Avestan punctuation characters. There are three empty unassigned slots covering the range 10B36–10B38. Eventually I decided to include these extra punctuation marks in the Avestan block. I do anticipate that a few more punctuation marks are to come and eventually there will not be enough room left in the Avestan block.

As discussed earlier in section 2.2, there is no fundamental notion of contextual shape change of characters in our proposal. Therefore, I see no need to add entries to `ArabicShaping.txt`. The joining behaviour of the final stems of the characters in Book Pahlavi is more similar to cursive variants of Latin than to Arabic.

⁴The character *ZWJ* can be used instead *Kashida* if desired.

	10BB	10BC	10BD
0			
1			
2			
3			
4			
5			
	10B3		
7		N/A	
8		N/A	N/A
9		N/A	N/A
A		N/A	N/A
B		N/A	N/A
C		N/A	N/A
D		N/A	N/A
E		N/A	N/A
F		N/A	N/A

Table 2.4: Proposed Book Pahlavi Character Mapping to Unicode along with two new punctuation characters. The punctuation characters are shared with Avestan

10B37		FOUR DOT DIAMOND PUNCTUATION
10B38		FOUR RING DIAMOND PUNCTUATION
10BB0		BOOK PAHLAVI LETTER BETH
10BB1		BOOK PAHLAVI LETTER ALTERNATE BETH-SIGN 1
10BB2		BOOK PAHLAVI LETTER YODTH
10BB3		BOOK PAHLAVI LETTER GIMEL-DALETH-YODTH-COMBINED SIGN 1
10BB4		BOOK PAHLAVI LETTER ALTERNATE GIMEL-DALETH-YODTH
10BB5		BOOK PAHLAVI LETTER ALTERNATE FINAL IH-ALTERNATE SIGN 1
10BB6		BOOK PAHLAVI LETTER OLD DALETH
10BB7		BOOK PAHLAVI LETTER WAW-NUN-AYIN-RESH
10BB8		BOOK PAHLAVI LETTER ZAYIN
10BB9		BOOK PAHLAVI LETTER KAPH
10BBA		BOOK PAHLAVI LETTER OLD KAPH
10BBB		BOOK PAHLAVI LETTER OLD LAMEDTH
10BBC		BOOK PAHLAVI LETTER LAMEDTH
10BBD		BOOK PAHLAVI LETTER L-LAMEDTH
10BBE		BOOK PAHLAVI LETTER MEM
10BBF		BOOK PAHLAVI LETTER PARTIAL SHIN
10BC0		BOOK PAHLAVI LETTER SAMEKH
10BC1		BOOK PAHLAVI LETTER SADHE
10BC2		BOOK PAHLAVI LETTER FINAL SADHE-PARTIAL PE
10BC3		BOOK PAHLAVI LETTER TAW
10BC4		BOOK PAHLAVI LETTER X1
10BC5		BOOK PAHLAVI LETTER X2
10BD0		BOOK PAHLAVI COMBINING DOT ABOVE
10BD1		BOOK PAHLAVI COMBINING TWO DOTS ABOVE
10BD2		BOOK PAHLAVI COMBINING THREE DOTS ABOVE
10BD3		BOOK PAHLAVI COMBINING CIRCUMFLEX
10BD4		BOOK PAHLAVI COMBINING DOT BELOW
10BD5		BOOK PAHLAVI COMBINING TWO DOTS BELOW
10BD6		BOOK PAHLAVI COMBINING THREE DOTS BELOW
10BD7		BOOK PAHLAVI COMBINING CARON BELOW

Table 2.5: Character names for the Unicode Standard


```

10B37;FOUR DOT DIAMOND PUNCTUATION;Po;0;ON;;;;;N;;;;;
10B38;FOUR RING DIAMOND PUNCTUAION;Po;0;ON;;;;;N;;;;;
10BB0;BOOK PAHLAVI LETTER BETH; Lo;0;R;;;;;
10BB1;BOOK PAHLAVI LETTER ALTERNATE BETH-SIGN 1; Lo;0; R;;;;;
10BB2;BOOK PAHLAVI LETTER YODTH; Lo;0; R;;;;;
10BB3;BOOK PAHLAVI LETTER GIMEL-DALETH-YODTH-COMBINED SIGND 1; Lo;0; R;;;;;
10BB4;BOOK PAHLAVI LETTER ALTERNATE GIMEL-DALETH-YODTH; Lo;0; R;;;;;
10BB5;BOOK PAHLAVI LETTER ALTERNATE FINAL IH-ALTERNATE SIGN 1; Lo;0; R;;;;;
10BB6;BOOK PAHLAVI LETTER OLD DALTEH-SIGN 10; Lo;0; R;;;;;
10BB7;BOOK PAHLAVI LETTER WAW-NUN-AYIN-RESH; Lo;0; R;;;;;
10BB8;BOOK PAHLAVI LETTER ZAYIN;Lo;0; R;;;;;
10BB9;BOOK PAHLAVI LETTER KAPH;Lo;0; R;;;;;
10BBA;BOOK PAHLAVI LETTER OLD KAPH;Lo;0; R;;;;;
10BBB;BOOK PAHLAVI LETTER OLD LAMEDTH; Lo;0; R;;;;;
10BBC;BOOK PAHLAVI LETTER LAMEDTH;Lo;0; R;;;;;
10BBD;BOOK PAHLAVI LETTER L-LAMEDTH;Lo;0; R;;;;;
10BBE;BOOK PAHLAVI LETTER MEM;Lo;0; R;;;;;
10BBF;BOOK PAHLAVI LETTER PARTIAL-SHIN;Lo;0; R;;;;;
10BC0;BOOK PAHLAVI LETTER SAMEKH;Lo;0; R;;;;;
10BC1;BOOK PAHLAVI LETTER SADHE;Lo;0; R;;;;;
10BC2;BOOK PAHLAVI LETTER FINAL SADHE-PARTIAL PE; Lo;0; R;;;;;
10BC3;BOOK PAHLAVI LETTER TAW; Lo;0; R;;;;;
10BC4;BOOK PAHLAVI LETTER X1;Lo;0; R;;;;;
10BC5;BOOK PAHLAVI LETTER X2;Lo; 0; R;;;;;
10BD0;BOOK PAHLAVI COMBINING DOT ABOVE; Mn;230; NSM;;;;;
10BD1;BOOK PAHLAVI COMBINING TWO DOTS ABOVE; Mn;230; NSM;;;;;
10BD2;BOOK PAHLAVI COMBINING THREE DOTS ABOVE; Mn;230; NSM;;;;;
10BD3;BOOK PAHLAVI COMBINING CIRCUMFLEX; Mn;230; NSM;;;;;
10BD4;BOOK PAHLAVI COMBINING DOT BELOW; Mn;220; NSM;;;;;
10BD5;BOOK PAHLAVI COMBINING TWO DOTS BELOW;Mn;220; NSM;;;;;
10BD6;BOOK PAHLAVI COMBINING THREE DOTS BELOW;Mn;220; NSM;;;;;
10BD7;BOOK PAHLAVI COMBINING CARON BELOW;Mn;220; NSM;;;;;

```

Table 2.6: The fragment of `UnicodeData.txt` pertaining to Book Pahlavi

2.11 Sorting

2.11.1 Collating basic characters

In modern scholarly books, the more common way of sorting Pahlavi characters is based on the Aramaic letter order. Essentially, the order of the Book Pahlavi characters is the same as the order of the Aramaic characters that have identical transliterations. Since a single Pahlavi character can usually be transliterated to multiple letters and correspond to any single one of the multiple Aramaic letters, the first Aramaic letter (in sorting order) is chosen. Therefore, the character 𐭪 which can be transliterated to any one of W, w, N, n, r , and etc. has the Aramaic sort key of W regardless of its transliteration. I remind the reader that sorting order for Aramaic letters is commonly taken to be as $A (^{\circ}) < B < G < D < H$ or $E < W < Z < \text{H} < \text{T} < Y < K < L < M < N < S < ^{\circ} < P < \text{š} < Q < R < \text{Š} < T$.

In most scholarly books the digraphs have sorting orders that are different from that of the combining characters (i.e., the constituent characters), because the digraphs are transliterated to a different Latin letter which has a different corresponding Aramaic letter. For example, the digraph 𐭪𐭫 —at least in isolation—is commonly transliterated to s , therefore it has the sorting order of Aramaic S , which means that it comes after the character 𐭫 which has transliteration m and hence sorting order of M .

Things get more complicated when, in practice, a given shape can be transliterated in multiple ways—say as a digraph versus two separate characters. In such cases *usually* the more frequent transliteration (often the digraph) is adopted and the order is determined regardless of the proper of the word. The exact choices may differ from source to source.⁵ As an example, the word 𐭪𐭫𐭬 which can be considered as $\text{𐭪} + \text{𐭫} + \text{𐭬}$ or considered as $\text{𐭪} + \text{𐭫𐭬}$, and transliterated as dyk' or sk' respectively has a single entry in say David Mackenzie's *A Concise Pahlavi Dictionary* [13]. The entry is after the words starting with 𐭫 (M), because $M < S$.

There are many more conventions to the collation rules. I have listed them in the table 2.7, based on [2, 13, 18]. The sources differ in a number of details, e.g., on how to break combinations of 𐭫 's and the sorting order of 𐭪 . I have normalized these differences to a single collation order listed in table 2.7. There are a few interesting points about table 2.7. First, note that the digraph 𐭫𐭬 with Aramaic transliteration as H or E is *not* considered a single unit when sorting is concerned and is taken to be $\text{𐭪} + \text{𐭬}$ (MN). This

⁵See the glossary in [18] as a counter example in which the transliteration and not the mere character combination is the basis.

Row	Shape	parts	Sort Key
1	𐬀	𐬀	B
2	𐬁	𐬁	B
3	𐬂	𐬀 + 𐬀 + 𐬀 + 𐬀	AGG
4	𐬃	𐬀 + 𐬀	A
5	𐬄	𐬀 + 𐬀	P
6	𐬅 ... 𐬅 ... 𐬅	𐬀 + N × 𐬀 + 𐬀	AA ^N
7	𐬆	𐬀 + 𐬀	AG
8	𐬇	𐬀	G
9	𐬈	𐬀	G
10	𐬉	𐬀 + 𐬀	S
11	𐬊	𐬀 + 𐬀	SG
12	𐬋	𐬀	D
13	𐬌	𐬀	DA
14	𐬍	𐬀	B
15	𐬎	𐬀	W
16	𐬏	𐬀	Z
17	𐬐	𐬀	K
18	𐬑	𐬀	K
19	𐬒	𐬀	L
20	𐬓	𐬀	L
21	𐬔	𐬀	Ł
22	𐬕	𐬀 + 𐬀	MW
23	𐬖	𐬀	M
24	𐬗 ... 𐬗 ... 𐬗	𐬀 + N × 𐬀 + 𐬀	DA ^N A
25	𐬘	𐬀 + 𐬀	DA
26	𐬙	𐬀	S
27	𐬚 ... 𐬚 ... 𐬚	𐬀 + N × 𐬀 + 𐬀	ŠA ^N
28	𐬛	𐬀 + 𐬀	Š
29	𐬜	𐬀	S
30	𐬝	𐬀	C
31	𐬞	𐬀	C
32	𐬟	𐬀	T
33	𐬠	𐬀	X1
34	𐬡	𐬀	X2

Table 2.7: Table of Pahlavi characters and digraphs and character combinations with different order. The transliterated system has collation order as A < B < G < D < W < Z < K < L < Ł < M < S < P < C < Š < T

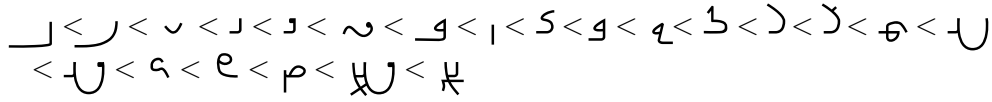


Figure 2.1: Sorting order of the basic characters with no digraph or combination

Pahlavi word	Comprising Characters	Sort Key	Applied rules
𐭪𐭫𐭬𐭭𐭮𐭯𐭰𐭱	[𐭪 + 𐭫 + 𐭬 + 𐭭]	ŠAA	29
𐭮𐭯𐭰𐭱	[𐭮 + 𐭯] + [𐭰 + 𐭱]	AP	4, 5
𐭮𐭯𐭰𐭱𐭲𐭳𐭴𐭵	[𐭮] + [𐭯 + 𐭰 + 𐭱 + 𐭲]	AAAT	6, 32
𐭮𐭯𐭰𐭱𐭲𐭳	[𐭮] + [𐭯] + [𐭰 + 𐭱]	SDC	10, 12, 30
𐭮𐭯𐭰𐭱	[𐭮] + [𐭯 + 𐭰]	DAM	25, 23

Table 2.8: Applying collation rules of table 2.7 to a list of words. The square brackets [] show the grouping of letters to which the rule is applied.

is a common convention among Pahlavi scholars. Also, in the case of rules 6, 23 and 26, N does not usually exceed 5; Nonetheless see figure 4.18 for an example of $N > 5$. Third, note the sorting order of the basic characters, which themselves follow the Aramaic order. I have noted them in isolation in figure 2.1 for emphasis.

When applying the rules of collation to determine the sorting order of a word, I start from row 1 of table 2.7 and proceed through rows until I hit a match for a sub-string starting from the character that I am at. At this point, I apply the corresponding sort key. Then, I start from the next character after the matched sub-string and then start from row 1 again.

Example

I want to sort the following words: 𐭪𐭫𐭬𐭭, 𐭮𐭯, 𐭮𐭯𐭰𐭱, 𐭮𐭯𐭰𐭱𐭲, 𐭮𐭯𐭰.

To sort them, I first break them to the basic characters and then I apply the collation rules and construct the sort keys. I have shown the process in table 2.8. Once I have the corresponding sort key for each word, it is easy to order them. From table 2.8 I can see that the collation order is 𐭮𐭯𐭰𐭱 < 𐭮𐭯 < 𐭮𐭯𐭰 < 𐭮𐭯𐭰𐭱𐭲 < 𐭪𐭫𐭬𐭭, corresponding to AAAT < AP < DAM < SDC < ŠAA. Remember that the Aramaic collation order of Latin characters has been used not the English order.

Having all that said, it is the author’s personal belief that such collation

rules (table 2.7) are somewhat unnecessary with such detail, *if* the current proposed encoding is used. Something like figure 2.1 would suffice. Since my encoding generates unique representations for Pahlavi words, it will eliminate the need for the arbitrary rules that were in place in sorting orders of *potential* multigraphs and ligatures; there would be no disagreement on sorting orders either. my proposal to simplify the sorting rules should cause no concerns. The collation rules of table 2.7 or similar are constructs put in place by orientologists to reduce the ambiguity and to simplify word look-ups. These rules are based on parallels that the *orientologists* have drawn between the Book Pahlavi script and the Aramaic script. The rules have little to do with how Pahlavi scribes would sort their own word-lists. Previous generations of orientologists have employed different collation rules. For example in [31] the Perso-Arabic ordering is used. The little extant material that I have on surviving Pahlavi (and Avestan) word lists and letter lists could be used to suggest different sort orders were employed. See [11] as an example; also see figure 4.3. The rules of collation in table 2.7 are also not ubiquitous enough to be implied without mention. Therefore, in the glossaries that I have consulted ([2, 13, 18]) the rules of collation are always explicitly stated to the reader.

2.11.2 Collating diacritics

The diacritics follow the sorting order of $\overset{\circ}{\circ} < \overset{\circ}{\circ} < \overset{\circ}{\circ} < \overset{\circ}{\circ} < \overset{\circ}{\circ} < \overset{\circ}{\circ} < \overset{\circ}{\circ} < \overset{\circ}{\circ}$. In particular, this order ensures that $\overset{\circ}{\circ}$ (transl. *g*) < $\overset{\circ}{\circ}$ (transl. *d*) < $\overset{\circ}{\circ}$ (transl. *y*). The diacritics have secondary weights as it is customary in the Unicode Standard.

2.12 Text standardization and normalization

The uses of diacritics, character separation, ligatures and variant characters differ from manuscript to manuscript. There are also corrupt usages and non-standard spellings, or even variations in spellings such as in the case of numbers. There may also be ambiguities in distinguishing some characters from similar-looking characters due to the handwritten nature of a manuscript.

In pedagogical and scholarly editions the text is edited and standardized. Spelling mistakes are fixed and so are the corrupt forms. In scholarly editions, the differences among the reference materials are of course clearly

documented. This editorial step of cleanup requires deep scholarly knowledge. The details of this process are beyond the scope of this document. I call this step *standardization*. Note that the output of this process can still contain diacritics and variant characters. In fact, the current proposal came out of the need to uniquely and unambiguously encode standardized Pahlavi texts, be it handwritten or typeset.

When running textual analysis on these standardized texts or perhaps across many pieces of text, some *normalization* (vs. *standardization*) needs to be performed as well to obtain best results. The goal of text normalization is to generate the *greatest common denominator* of different variations or versions of a standard (scholarly edited) piece of writing and to arrive at the canonical form. The normalization can be automated and does not require scholarly knowledge.

A given Pahlavi word encoded according to my proposal can be normalized to its canonical form by applying the following transformations:

- Removal of ZWNJ and non-breaking hair spaces and tatweels
- Removal of diacritics
- Substituting all the variant (alternate) characters (the second column in table 2.1) with their corresponding more common variant.

Of course depending on the situation partial normalization, e.g. performing a subset of these steps or getting rid of less common diacritics, can be performed as well.

Chapter 3

The Problems with the Previous Encoding Models

There are three different classes of problems associated with the encoding models of the previous proposals ([7, 20]).

- Non-unique encoding to Unicode
- Non-unique rendering from Unicode
- Requiring complex rendering and obligate ligatures at the font level and render engine level, and possibly requiring high-level markup for accurate representation and preventing the loss of information.

Before moving forward, I must acknowledge that the existence of the first two class of problems has been postulated by the author of [20]. However being mere postulations, no examples were provided by the author.

As for the first class of problems (non-unique encoding) the proposal [20] itself has correctly conjectured that following its model “there may still exist several different ways to represent in Unicode a piece of writing on paper” [20, p.3]. The author then asserts that “[if such a case exists it would be] unavoidable considering all the ambiguities of Book Pahlavi” [20, p.3]. In section 3.1 I will show multiple examples for which multiple encodings exist for a single piece of writing on paper according to [20]. I will also show that all those examples have *unique* encodings according to our model and all those problems are avoided.

As for the second class of problems (non-unique decoding and rendering), the author of [20] has put forward the concern that “[s]ometimes more

than one ligature form is available for a combination of two or more characters” [20, sec.18, p.8]. This is indeed a manifestation of the second class of problems. He has assumed that in such cases both alternatives would be acceptable [20, sec 18, p.8]. In section 3.2 I provide examples of such problems and also show that the different renderings contain non-overlapping information, therefore choosing one instead of the other will have real-world consequences. The author of [20] does put forward remedies in case his assumption is wrong. The remedies are in the form of adding new ligature characters to the standard or adding higher-level markup or similar. His potential remedies bring us to the third class of problems.

In general, it seems that [20] considers the third class of problems (requiring extra markup and lots of obligate ligatures) an unavoidable necessity. my goal is to show that this class of problems is both avoidable and unnecessary.

Below, I discuss the first two classes of problems in detail and provide examples. I will also see that—in the context of the examples provided—the third class of problems is avoided and unnecessary when following my method of encoding.

3.1 Encoding to Unicode

Any encoding system of Book Pahlavi that proposes a mapping in which the **domain** is the all the Latin transliteration letters used by the scholarly community, and the **range** is a set where each member of the set is a *single* Unicode character, will yield multiple (degenerate) solutions in some cases. Simply put, going by such mapping, there exist cases in which a single word may be validly encoded in two or more completely different strings. Unfortunately, all the previous proposals suffer from this shortcoming, as I will show shortly. This is partly because they have chosen to encode digraphs as separate Unicode characters.¹ To be more specific, note that the proposal [20] has proposed a separate character for each of the elements in table 2.2.

In the case of Pahlavi, degenerate solutions can both manifest at the character level and at the word level, if I adopt the encodings put forward by the previous proposals.

¹As an analogy the problem with the encoding methods of the previous proposals is like that of defining a new Unicode character “sh” (which is simply “s” + “h”), that is supposed to represent the š phoneme in English. In such a case the simple word *glasshouse* can be written either as ...+s+s+h+o... or ...+s+sh+o+.... Needless to say that encodings like this will only lead to confusion and fragmentation.

As for a simple case, consider the Pahlavi word 𐭥𐭥 . As mentioned earlier, it can be transliterated either as *dyk'* standing for the word *dēg* meaning cauldron or transliterated as *sg'*, standing for the word *sag*, meaning dog [13]. Going by the encoding method of the previous proposal, the fragment 𐭥𐭥 can be encoded as $\text{𐭥} + \text{𐭥}$ ($\langle 10\text{BBD}, 10\text{BB7} \rangle_p$ according to [20])² or $\text{𐭥} + \text{𐭥} + \text{𐭥}$ ($\langle 10\text{BB2}, 10\text{BB2}, 10\text{BB7} \rangle_p$ according to [20]), with one case being transliterated as *sg* and the other one as *dyk*. Even with a knowledge of Middle Persian (to reduce the set of possible solutions) the proposed system in [20] cannot uniquely encode the word: $\langle 10\text{BBD}, 10\text{BB7} \rangle_p$ vs $\langle 10\text{BB2}, 10\text{BB2}, 10\text{BB7} \rangle_p$.

In my system, however, there is only one way to encode the word: $\text{𐭥} + \text{𐭥} + \text{𐭥} + \text{𐭥}$ ($\langle 10\text{BB4}, 10\text{BB4}, 10\text{BB9}, 10\text{BB7} \rangle_m$). Then the scholar can decide how to read it based on the context, the typist need not know Pahlavi.

For a more complex example that involves the use of multiple different characters (and obligate ligatures if I go by the previous proposals) consider the word 𐭥𐭥𐭥 . It can be transliterated as *gy^hh*, standing for *giyā(h)* (meaning grass) or transliterating as *syd^h*, standing for *syā* (meaning black). If I proceed to encode the word into Unicode using a system like that of [20], I will arrive at two completely different encoded strings based on my starting choice of the transliteration.

If I go by *gy^hh* then the proposal [20] would prescribe $\text{𐭥} + \text{𐭥} + \text{𐭥} + \text{𐭥}$ ($\langle 10\text{BB2}, 10\text{BB2}, 10\text{BB0}, 10\text{BB0} \rangle_p$ in [20]). Moreover, in order to produce the correct shape of the word, I have to use the ligature rules that $\text{𐭥𐭥} \leftarrow \text{𐭥} + \text{𐭥}$ and the rule that $\text{𐭥𐭥} \leftarrow \text{𐭥} + \text{𐭥}$. None of these rules are captured at the Unicode level. They need to be either handled at the layout enging, font level or some high-level markup method; none is desirable.

Now if I go by *syd^h*, then according to the encoding system put forward in [20] I need to encode the word as $\text{𐭥} + \text{𐭥} + (\text{𐭥}) \text{𐭥} + \text{𐭥}$ ($\langle 10\text{BBD}, 10\text{BB2}, 10\text{BB2}, 10\text{BB0} \rangle_p$ in [20]). This time in order to produce the correct shape of the word first I need choose the alternate glyph 𐭥 of 𐭥 ($\langle 10\text{BB2} \rangle_p$ in [20]). Then I have to use the ligature rules $\text{𐭥𐭥} \leftarrow \text{𐭥} + \text{𐭥}$ and $\text{𐭥𐭥} + \dots \leftarrow \text{𐭥} + \text{𐭥} + \dots$. Again, these rules are not captured at the Unicode level. And of course, $\langle 10\text{BBD}, 10\text{BB2}, 10\text{BB2}, 10\text{BB0} \rangle_p \neq \langle 10\text{BB2}, 10\text{BB2}, 10\text{BB0}, 10\text{BB0} \rangle_p$.

The implications of this shortcoming are sever. As an example, if Alice

²Throughout this chapter I use the index p when referring to the encoding method put forward in [20] and use the index m when using the character encoding model put forward by us.

as a student of Middle Persian encounters a word that she does not know how to transliterate, she would have a very hard time searching for the word say on a simple dictionary website. There are a few ways for her to find the meaning of the word, all of them being non-desirable. She can try all the encoding combinations, provided that she is even familiar with properly transliterating the ligatures (as ligatures are needed in the previous proposals). Or perhaps the website that she is searching on has some rules to simplify the search process, e.g. **𐭥** is always encoded as **𐭥𐭥** and never as **𐭥 + 𐭥** even if it is transliterated as *gg*— similar in spirit to the conventions in of collation rules. This will certainly reduce her search space but she still has to have some external information, and has to work much harder than necessary.

In my proposed encoding method, however, the word can be uniquely and unambiguously encoded as **𐭥 + 𐭥𐭥 + 𐭥𐭥𐭥 + 𐭥** (<10BB4, 10BC0, 10BBF, 10BB3>_m). No complicated ligature rules or high-level markups are needed either. In fact problems like these were my constraints around which I developed the proposed encoding method.

As demonstrated so far, it is a matter of great importance to separate linguistic and transliteration matters from that of canonical native Pahlavi text representation. A good encoding method should present a complete and *irreducible* set of characters that are able to uniquely and unambiguously encode a given Book Pahlavi text passage. It is then a matter of scholarly expertise and *opinion* to transliterate the encoded text into the Latin alphabet.

Perhaps, partially due to a lack of encoding standard, most scholarly books have chosen to reduce the Pahlavi script to a set of glyphs that map well with their transliteration schemes. This means that the mental basis for representing texts is not just the atomic elements of the script itself but rather the basis is mixed with linguistic elements and the historical baggage of the script that pollute the entire basis and make it needlessly complex. In this case, not only words cannot be unambiguously represented, but also complex ligature rules have to be introduced.

3.2 Decoding and rendering from Unicode

So far I have shown that if I adopt the encoding methods put forward in the previous proposals, in some cases, even for some simple Pahlavi words, multiple encoding solutions can be generated. The reverse of the statement is also true. It means that starting from an existing text fragment that

is encoded to the Unicode characters put forward by one of the previous proposals, multiple words can be generated. In other words, using those encodings, sometimes the original word cannot be uniquely recovered and hence there would be some loss of information with potentially severe real-world effects.

As an example assume that Alice is entering a Pahlavi text from an old manuscript into her institute’s database using the encoding standard put forward by [20]. Now assume that she encounters the fragment 𐭪𐭫 . By her past experience, and now being familiar with Pahlavi ligature rules that are a necessity in [20], she correctly identifies the shape as a ligature transliterating to “ls”. Lucky for her, according to the standard that she is using to encode, there is only one way to encode “ls” and that is $\text{𐭪} + \text{𐭫}$ ($\langle 10BB9, 10BBD \rangle_p$ in [20]).

Now assume that Bob downloads and prints out the text that was entered by Alice, as Bob wants to make a fresh translation of the passage entered by Alice.

Without any extra high-level markup or variant glyphs, and just by following the standard as proposed by [20] in plain text, the rendering engine would represent that “ls” fragment as 𐭪𐭫 . The reason is that $\text{𐭪} \leftarrow \text{𐭪𐭫}$ is not always exercised, i.e. 𐭪𐭫 is a *valid* form [2]. Now Bob is in some trouble, because he needs to consider all the possible transliterations of 𐭪𐭫 e.g., *lgy*, *lgy*, *ls* to name a few. Had he been presented with the original shape, he would have no trouble transliterating it as *ls* as Alice had done so.

So in order to prevent the loss of information in the encoding system [20], some meta information needs to be added by Alice, e.g. the ligature should be encoded as a separate character, or perhaps using variant glyphs for the font. This is unnecessary and extra burden on all parties, font developers, typesetters and other involved parties.

Again, had Alice been using my proposed encoding system, she would simply encode the word as $\text{𐭪} + \text{𐭫}$ ($\langle 10BBC, 10BBF \rangle_m$) and Bob would have it printed as 𐭪𐭫 as well, so there would be no loss of information.

To give an example with a complete word, consider the word $\bar{a}sm\bar{a}n$, transliterated as ${}^s m {}^n$, meaning the sky. It can be written as either 𐭪𐭫𐭬 or 𐭪𐭫𐭬 both being valid forms [31, p.26]. The proposal [20] encodes both forms as $\langle 10BB0, 10BBD, 10BBC, 10BB0, 10BB5 \rangle_p$. Going through the round trip, following [20], I may start with 𐭪𐭫𐭬 and I will end up with 𐭪𐭫𐭬 . The second form has more ambiguity due to the existence of multiple different readings for 𐭪 . Therefore, during the round-

trip some information has been lost. Similar to the previous example, Alice needs to somehow encode the alternate glyph for the phoneme *s*. This is not easily done in plain text. Of course in my proposed model the first form of the word is encoded uniquely as $\text{𐭠} + \text{𐭡} + \text{𐭡} + \text{𐭢} + \text{𐭣} + \text{𐭤} + \text{𐭥} + \text{𐭦}$ (<10BB3, 10BB3, 10BC0, 10BBE, 10BB3, 10BB3, 10BB7>_m), and the second form of the word is encoded uniquely as $\text{𐭠} + \text{𐭡} + \text{𐭡} + \text{𐭢} + \text{𐭣} + \text{𐭤} + \text{𐭥} + \text{𐭦}$ (<10BB3, 10BB3, 10BB4, 10BB4, 10BBE, 10BB3, 10BB3, 10BB7>_m).

In some cases, such as in the Pahlavi word 𐭣 (*ws*) transcribed as *was* meaning *much* or *many* or *enough*, the form constructed based on the model proposed in [20] ($\text{𐭣} \leftarrow \text{𐭤} + \text{𐭥}$) is *not* the correct spelling, and therefore is not an acceptable alternative; see [31, p.153] and [13, p.190]. This example serves to counter the conjecture in [20, sec.18, p.8] regarding acceptability of the alternate form.

Chapter 4

Sources and Examples

4.1 Samples from different sources

The Pahlavi sources that were consulted for writing this proposal fall under four categories:

- Pre-nineteenth century hand-written manuscripts written in Iran and India. I have used high-quality digital versions of photographs or facsimiles [3, 4, 24, 28, 29].
- High-quality Pahlavi texts and glossaries published in India in late nineteenth and early twentieth century, by Parsee Indians and sometimes Western scholars [3, 6, 8–12, 14, 23, 31].
- High-quality type-set passages and fragments using modern typographic techniques and modern equipment and computers [1, 15, 17, 18, 21].
- Handwritten passages by modern scholars or scribes when compiling assorted material, or dictionaries or other scholarly or religious documents [2, 13, 16, 19, 21].

Below I will provide pictures of pages from sources in each category for elucidation and point out some interesting features. Finally I will encode a sample from a source for illustration.

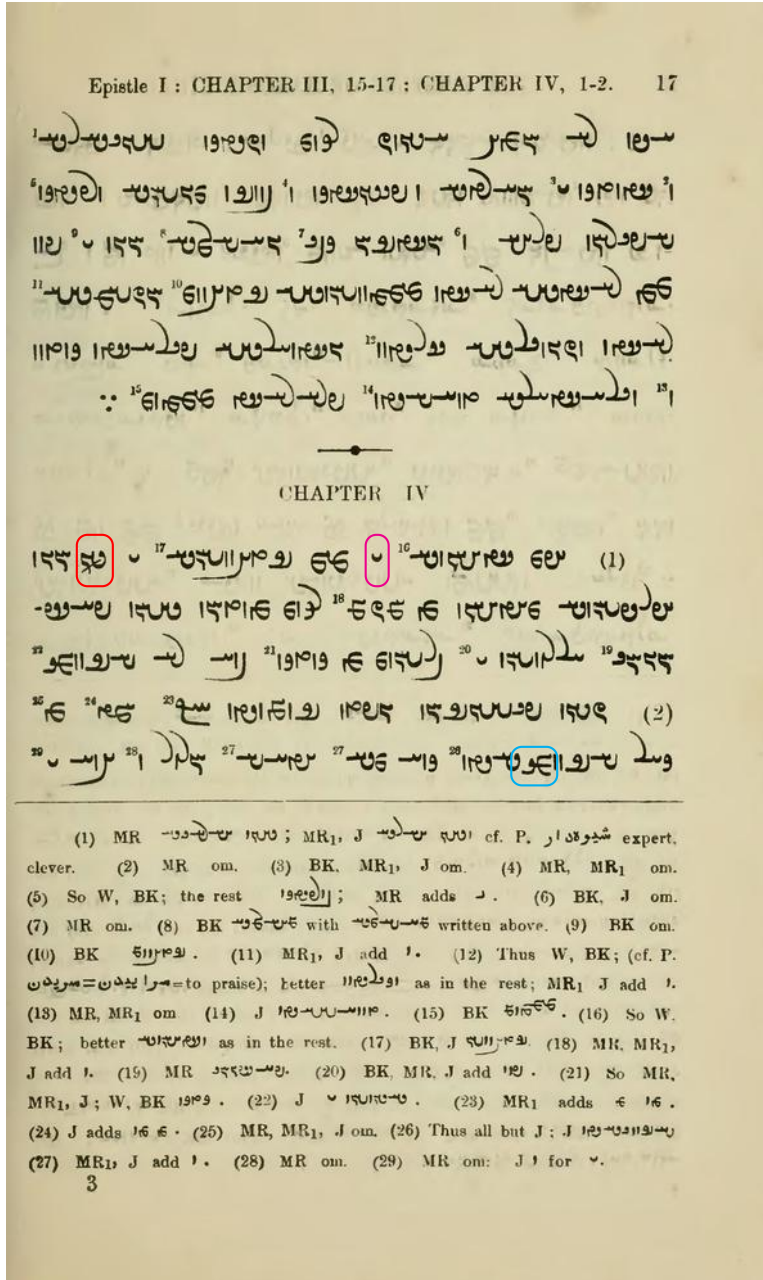


Figure 4.1: A page from [6]. Note the use of characters 𐭆𐭇 (red) and 𐭆𐭈 (cyan). Also note the genitive preposition 𐭆𐭇𐭈 (magenta).

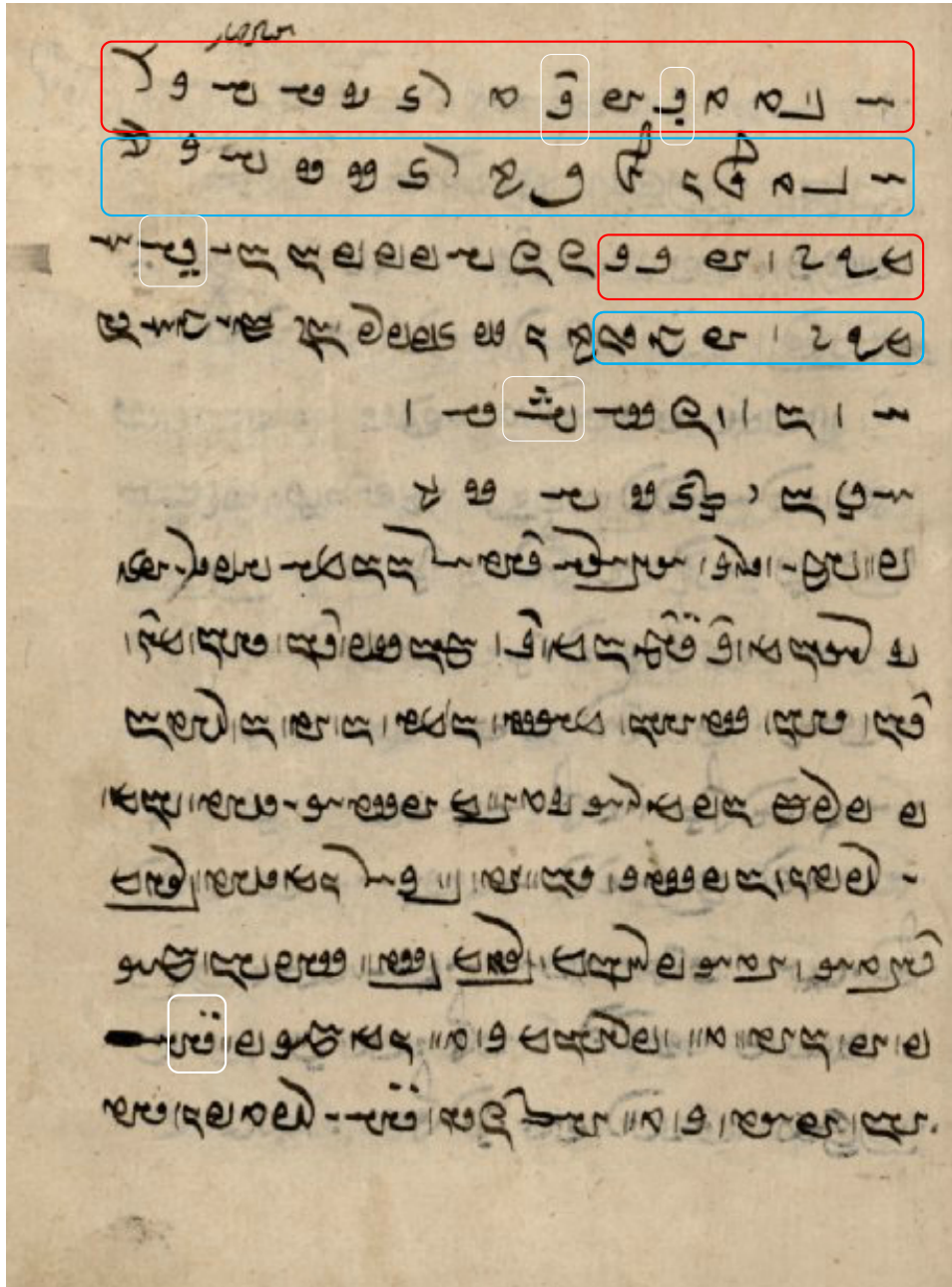


Figure 4.3: A page from a Pahlavi manuscript [27, p.166]. The section marked red lists the primarily Pahlavi letters (mixed with three Avestan letters) in approximately the Perso-Arabic order. The cyan section below each red section lists the equivalent fully Avestan script version. The Pahlavi letters from right to left are: **line 1:** *a, b, t, θ, j, x* (Avestan *h*), *d, δ, r, z, s, s, š, k, l*, **line 2:** *m, γ, n̄* (Avestan), *n, h* (Avestan), *y, g*. The transcription of the Pahlavi letters was made with reference to their listed Avestan equivalent below. Note the diacritics ̇, ̈, ̉, ̊, ̋ (white).

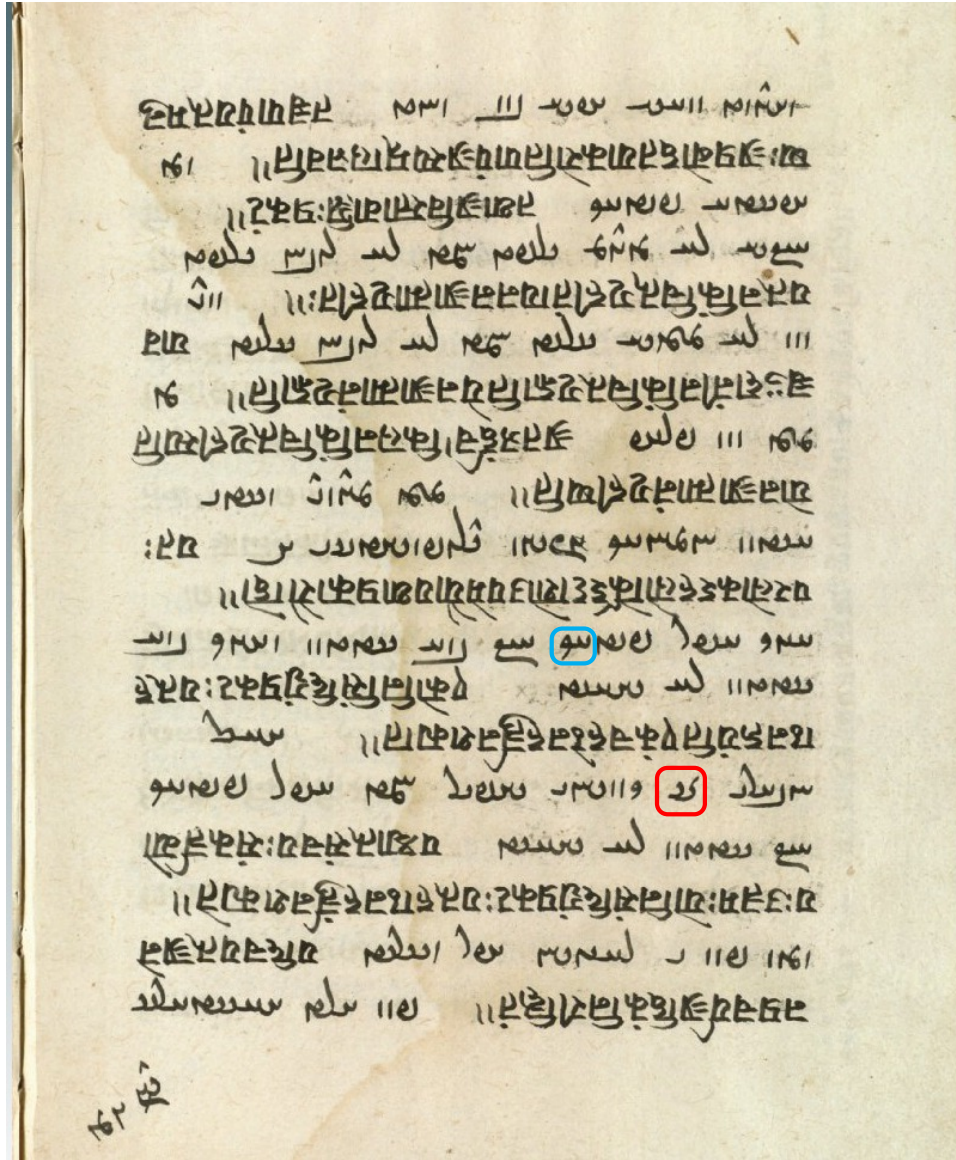


Figure 4.4: A page from [25]. Note the corrupt form of ज as ज in the word ज in line 15 (red). Normal forms of ज can also be seen, e.g. line 12 (cyan). The other language present also in the text is Sanskrit.

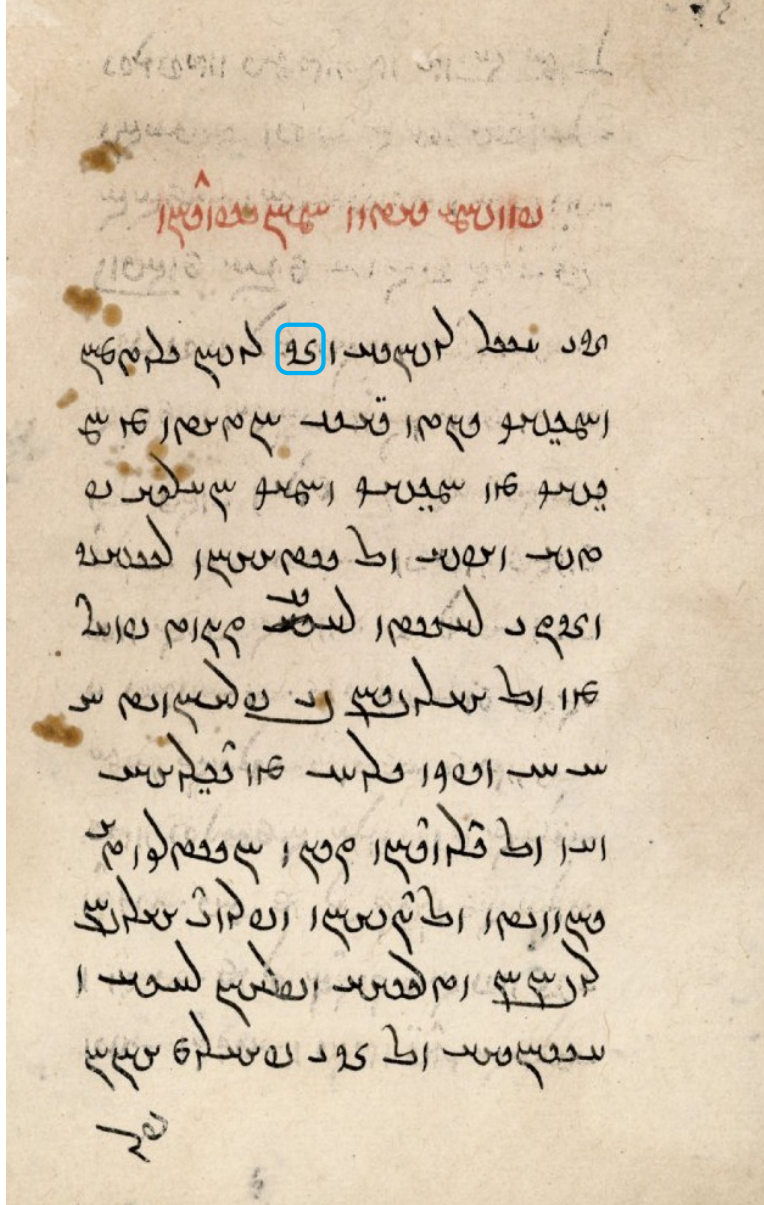


Figure 4.5: The first page of [29]. Note the conventional form (as opposed to the corrupt form) of the $ق$ in $قو$ in line 2 (cyan). Also note the vertical kerning of $ق$ in the word.

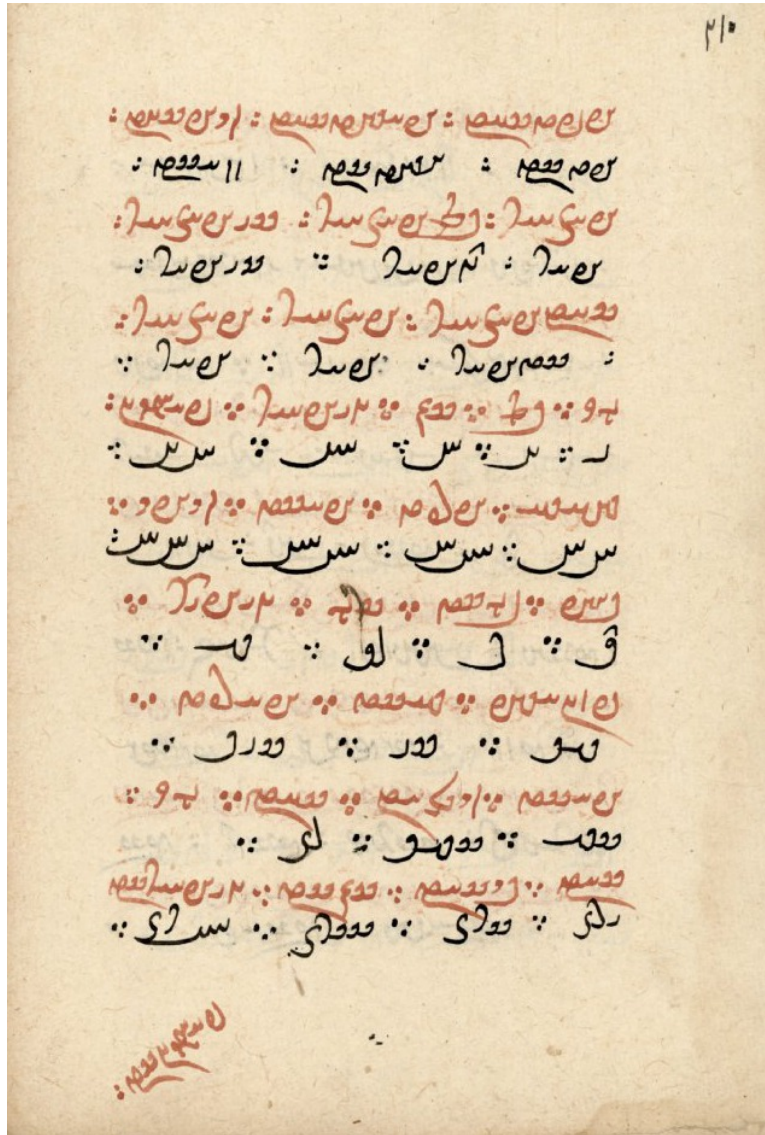


Figure 4.7: A page of Pahlavi word list in [26, f.210b]. The red text is Pāzand, and the black text below it, is the Pahlavi form. The Pahlavi numbers start from the fourth black line (line 8). **Line 8** reads: 1, 2, 3, 4, 5. **Line 10**: 6, 7, 8, 9. **Line 10**: 10, 20, 30, 40. **Line 12**: 50, 60, 70. **Line 14**: 80, 90, 100. **Line 16**: 1 (×) 100, 200, 300, 400. Note the use of 𐭪 . Finally, note that even in case of 50 the character for 10 is different than 𐭪 , although not as elongated as 10 for example. However, in figure 4.8, the 𐭪 in 50,000 is elongated beyond doubt. This confirms my assertion for 50.

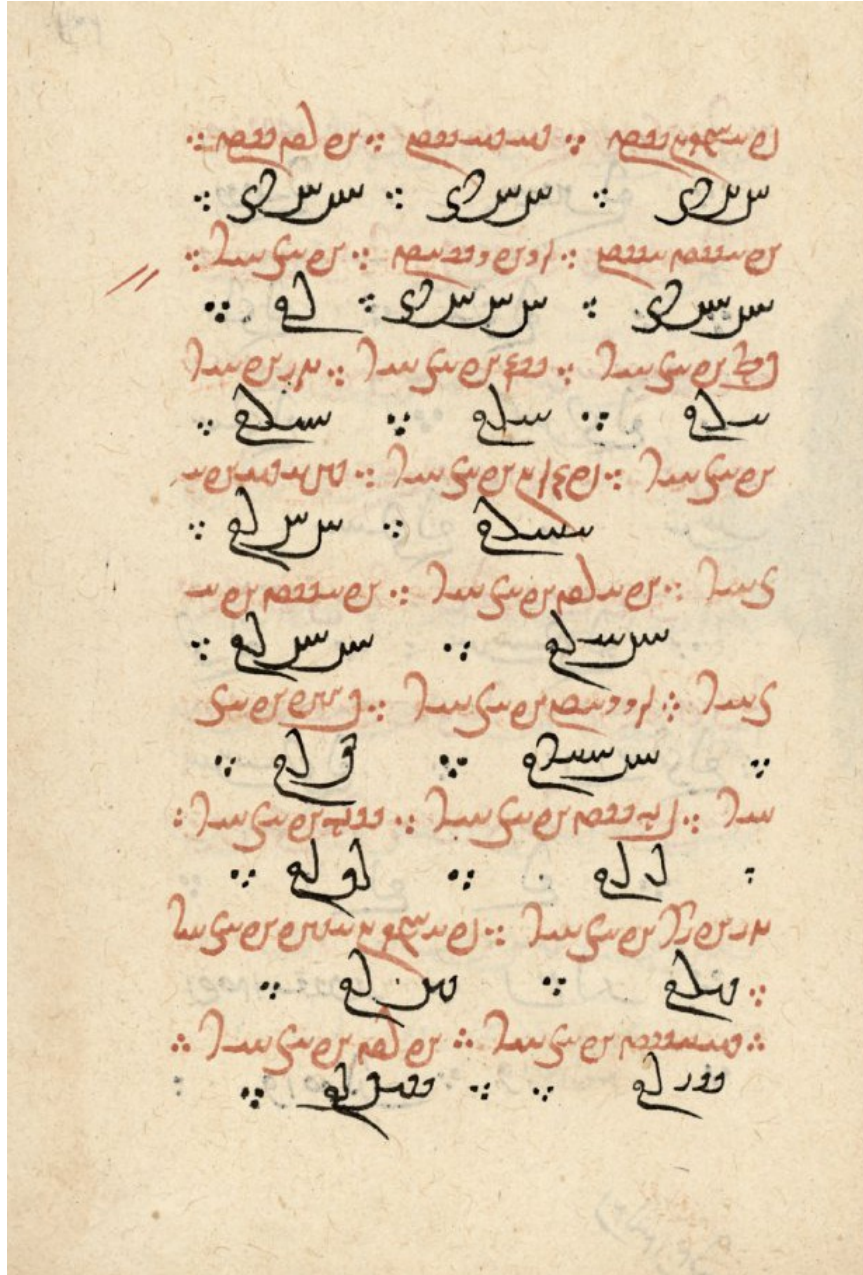


Figure 4.8: A fragment of Pahlavi and Pāzand text from [26, f.211a] that depicts of assorted numbers from 500 to 70000. **Line 2:** 500, 600, 700. **Line 4:** 800, 900, 1000. **Line 6:** 2000, 3000, 4000. **Line 8:** 5000, 6000. **Line 10:** 7000, 8000. **Line 12:** 9000, 10000. **Line 14:** 20000, 30000, **Line 16:** 40000, 50000. **Line 18:** 60000, 70000. Note the use of 𐭌 in lines 14 and 12

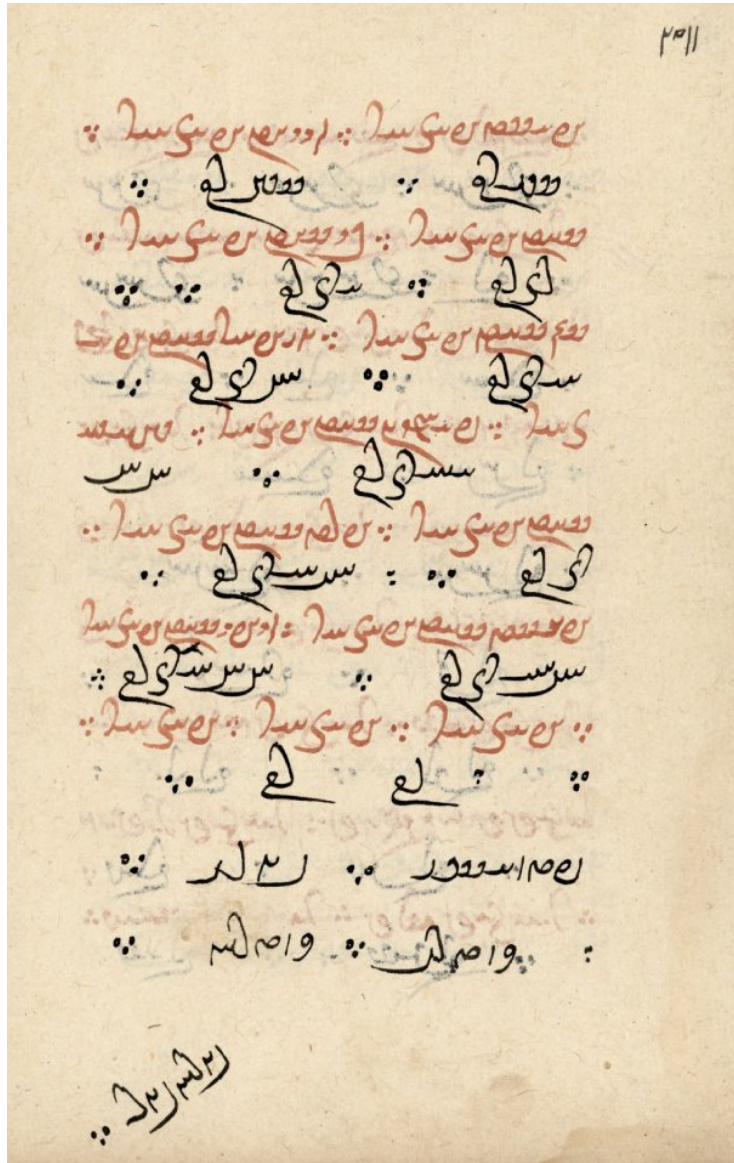


Figure 4.9: A fragment of Pahlavi and Pāzand text from [26, f.211b] that depicts the assorted numbers from 80,000 to 1,000,000. **Line 2:** 80000, 90000. **Line 4:** 100,000, 200,000. **Line 6:** 300,000, 400,000. **Line 8 and 10:** 500,000, 600,000, 700,000. **Line 12:** 800,000, 900,000. **Line 14:** $1000(\times)1000 = 1,000,000$. Note the line break in the middle of 600,000, between *سرس* and *کله*. However we do not recommend such behaviour for the standard and this case should be considered special circumstance.

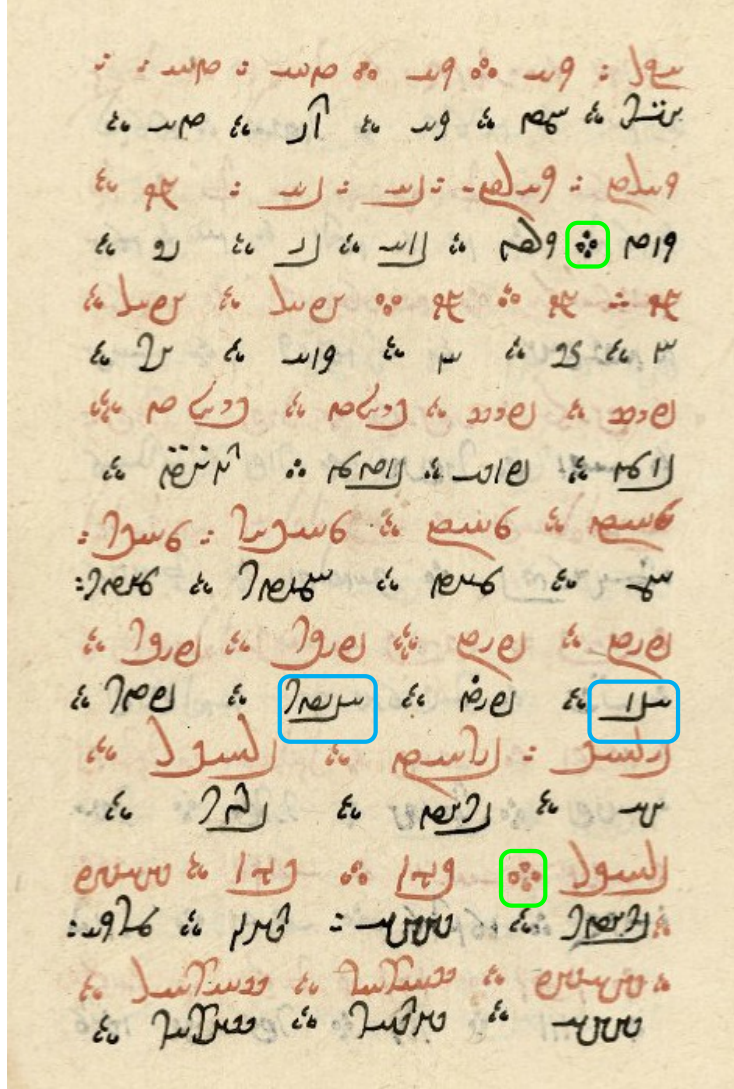


Figure 4.10: Another page of Pahlavi word list in [26, f.198a]. Note the first and third word in line 12, containing the combination س . Note the angularity of س and compare the whole combination to (most) instances of س in the same word list in the manuscript shown in figures 4.7 and 4.8. Also note the separation character composed of four large dots (or rings), line 4 from the top and line 4 for the bottom.

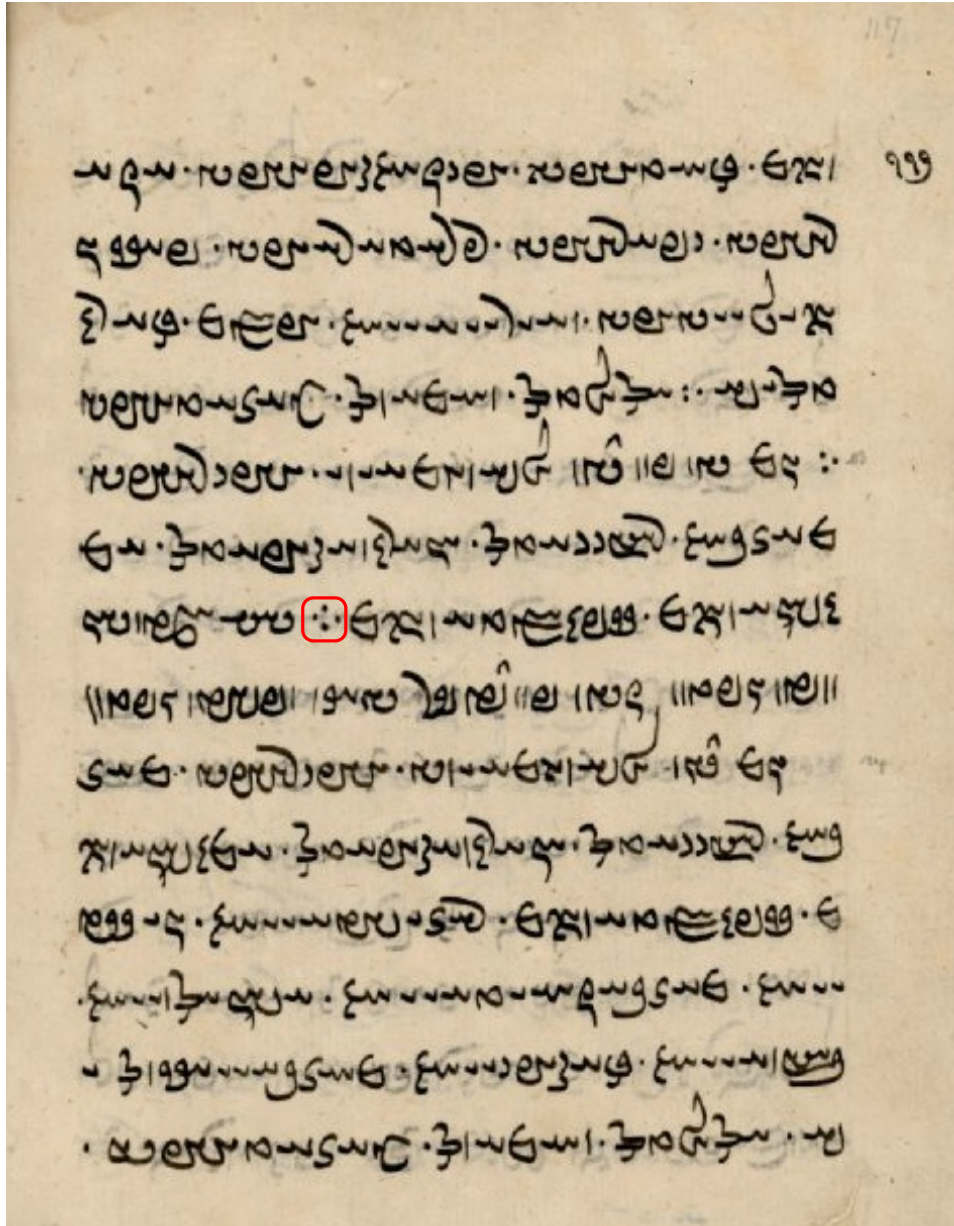


Figure 4.11: A page from a Pahlavi/Avestan manuscript [27]. The 4-dot punctuation mark is boxed in red. Most of the page is written in Avestan script.

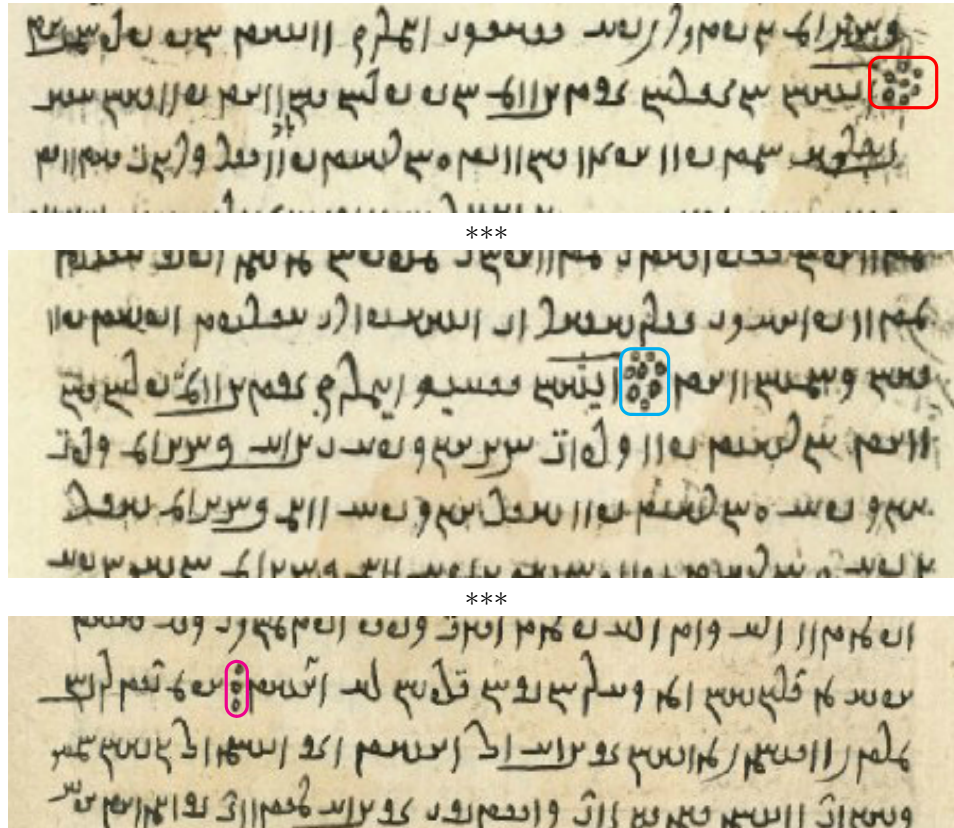


Figure 4.12: Fragments of a manuscript [28] showing punctuation marks that are not proposed for encoding yet. The first fragment is taken from page 72, the second from page 71 and the third from page 42.

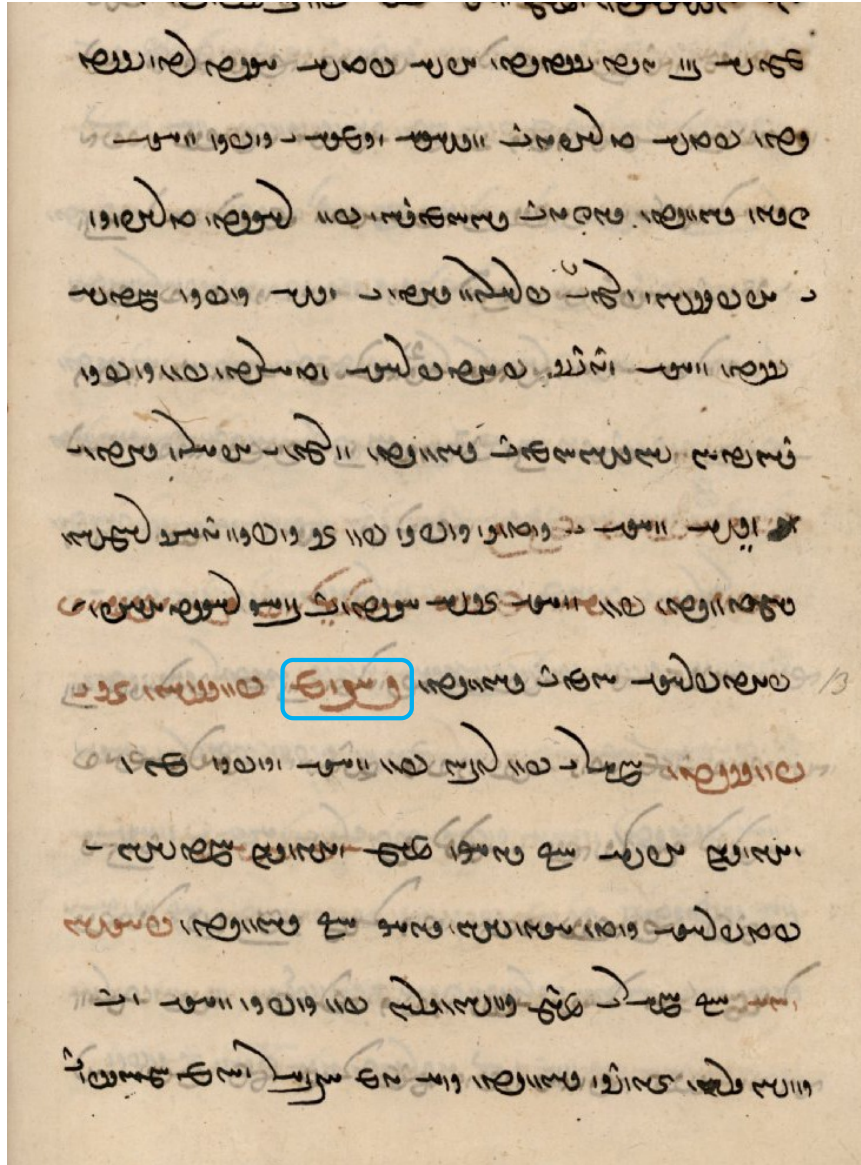


Figure 4.13: A fragment from manuscript [24]. Note the ordinal number 13th (13^m) in red written as **ع** **س** **و** on line sixth from the bottom (cyan).

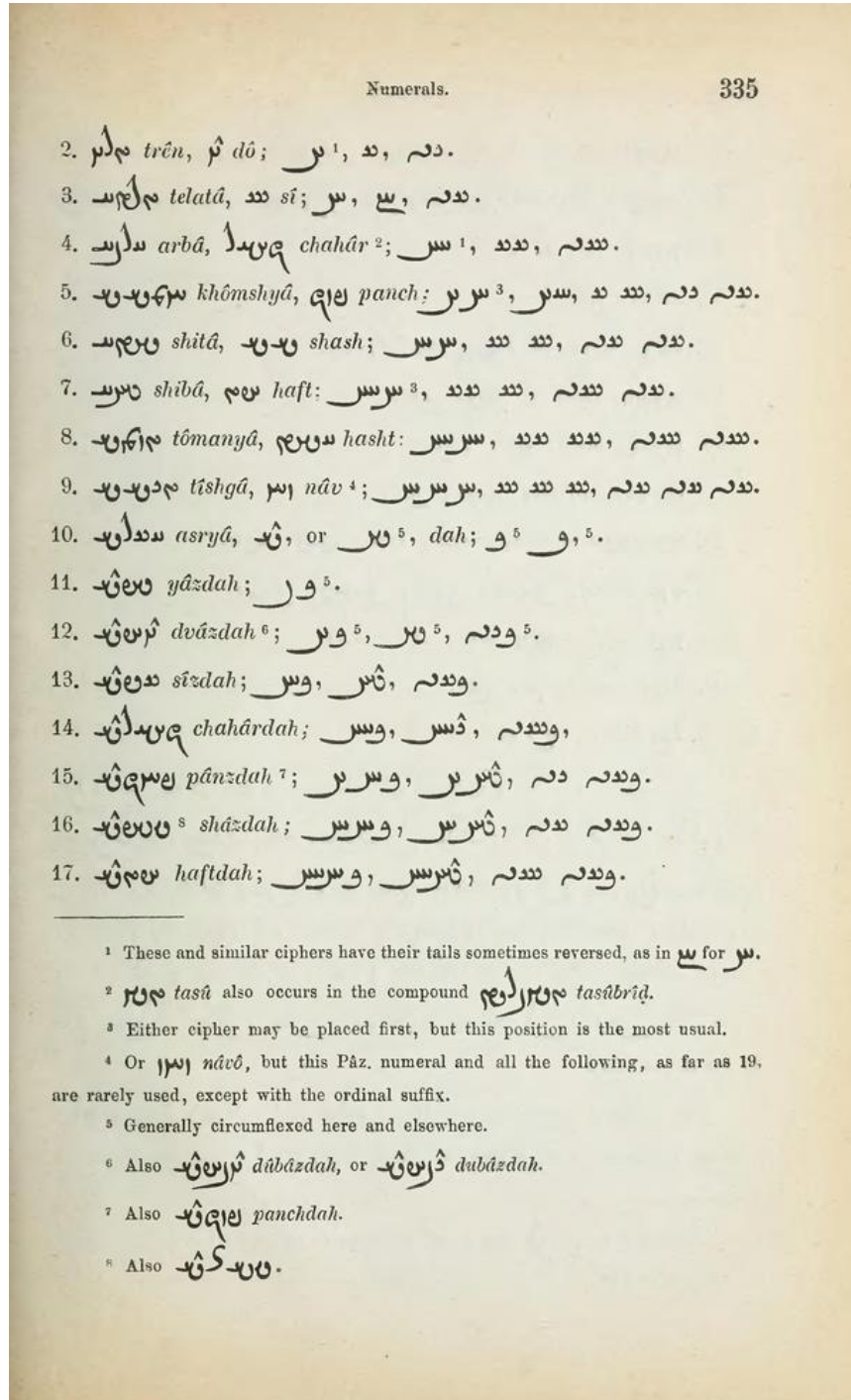


Figure 4.14: Variants of numerals from 2–17 using different ciphers (taken from [31]).

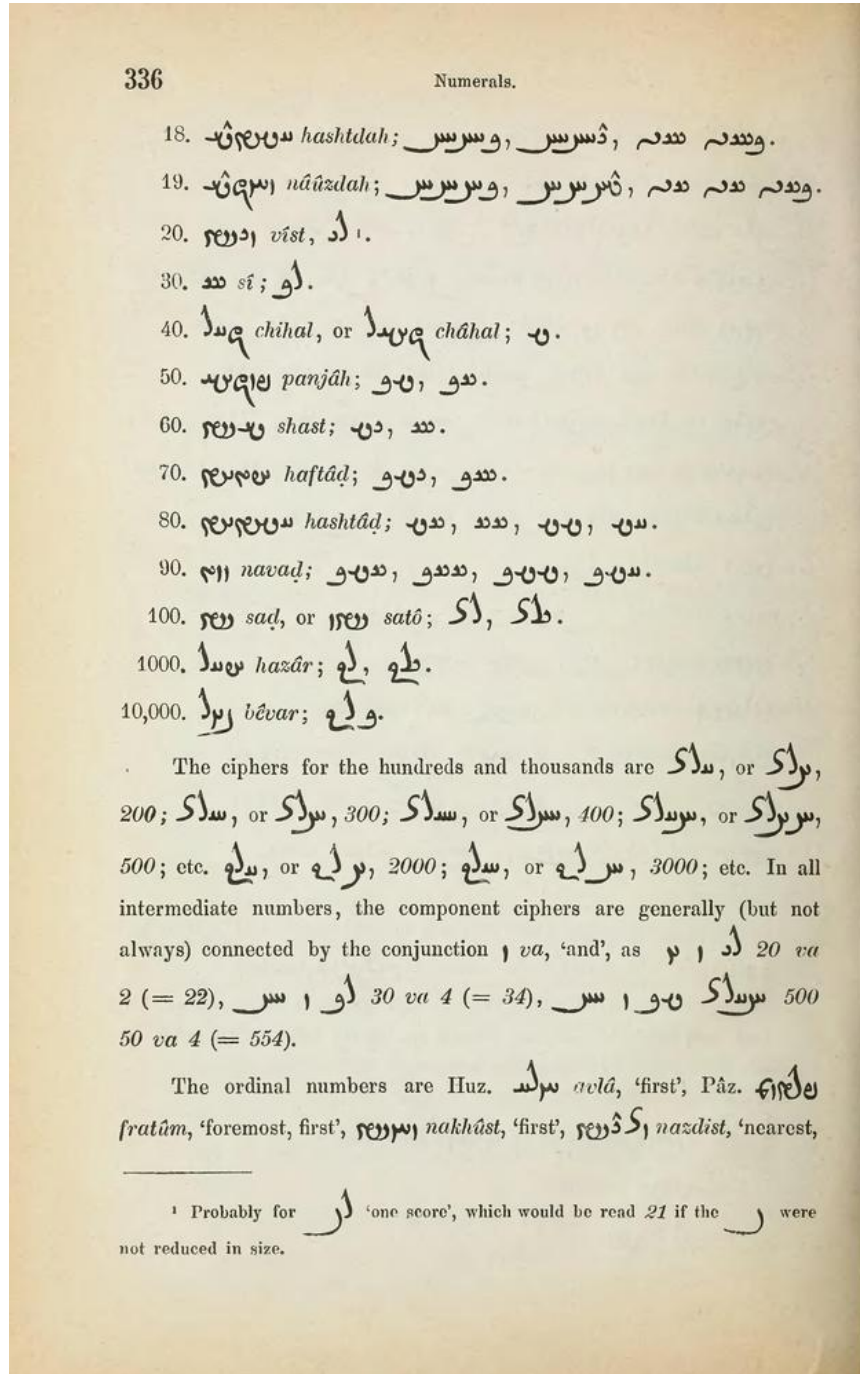


Figure 4.15: Variants of selected numerals from 18–10,000 using different ciphers (taken from [31]).

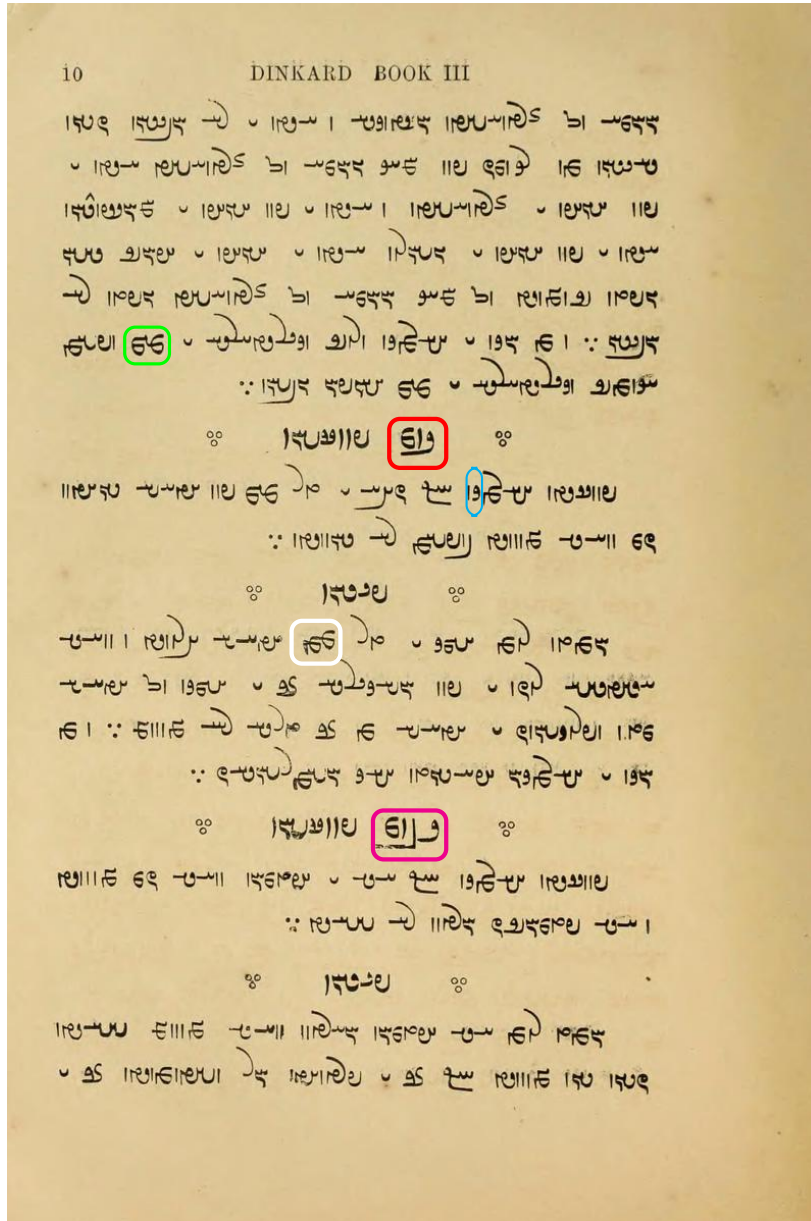


Figure 4.17: A page from Book III of *Dinkard* [14]. Note the ordinal number 10th in line 8, 𐬨𐬀 (red). Compare the character 𐬨 in the word with an isolated 𐬨, e.g., line 9 second word (cyan), which shows that they are different characters. Also note that the hat is dropped. Number 11, i.e. 𐬨𐬀 can be seen in line 16 (magenta). Note that this time the simple character 𐬨 is used instead of the variant 𐬨. Finally note 𐬨𐬀 (green) where both 𐬨's are more or less on the baseline, but not so in 𐬨𐬀 (white).

.. ١٤٤٤ " [١٤٤٤] ١٤٤٤ [١٤٤٤] ١٤٤٤
 ١٤٤٤ ١٤٤٤ [١٤٤٤] ١٤٤٤ ١٤٤٤ ١٤٤٤.

١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤
 ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤
 ١٤٤٤ [١٤٤٤] ١٤٤٤ [١٤٤٤] ١٤٤٤
 ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤
 ١٤٤٤ [١٤٤٤] ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤
 ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤
 ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤ ١٤٤٤

Figure 4.20: A fragment from *Zand of Srōš Yasn* taken from [21]. The typeface is similar to that of figure 4.19. However, note the difference that there is vertical kerning of 𐬐 in the word 𐬕𐬀 (red) with the same word in figure 4.19. Also note the word space between numbers on line 5: 𐬕𐬀 𐬕𐬀 𐬕𐬀 (cyan). Finally note the use of corrupt form of 𐬐 in number 50, where 𐬐 is used instead of 𐬐 in 𐬕𐬀 (magenta).

مترما ۹۱

مینوی خرد

پرسش ۲

۱ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۲ سے لہاں۔
 ۲ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۳ سے لہاں۔
 ۳ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۴ سے لہاں۔
 ۴ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۵ سے لہاں۔
 ۵ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۶ سے لہاں۔
 ۶ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۷ سے لہاں۔
 ۷ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۸ سے لہاں۔
 ۸ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۹ سے لہاں۔

۳

۱ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۲ سے لہاں۔

۲ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۳ سے لہاں۔
 ۳ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۴ سے لہاں۔
 ۴ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۵ سے لہاں۔
 ۵ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۶ سے لہاں۔
 ۶ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۷ سے لہاں۔
 ۷ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۸ سے لہاں۔
 ۸ رهندها و سوا اک ۶۱۲۰۰ ر ۱۱۱۱ ۹ سے لہاں۔

Figure 4.21: A passage from *Minug i Xrad* (starting from the second question), taken from [2]. The passage is handwritten in a fairly clear pedagogical style, making it easy for novices to read the text. At the same time the typeface of the handwriting employs some of the common stylistic ligatures. In section 4.2 we will encode the red boxed section.

۱۶ ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ ۲۸ ۲۹ ۳۰ ۳۱ ۳۲ ۳۳ ۳۴ ۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰ ۴۱ ۴۲ ۴۳ ۴۴ ۴۵ ۴۶ ۴۷ ۴۸ ۴۹ ۵۰ ۵۱
 Ankl.13 | ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ ۲۸ ۲۹ ۳۰ ۳۱ ۳۲ ۳۳ ۳۴ ۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰ ۴۱ ۴۲ ۴۳ ۴۴ ۴۵ ۴۶ ۴۷ ۴۸ ۴۹ ۵۰ ۵۱
 DH.114 | ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ ۲۸ ۲۹ ۳۰ ۳۱ ۳۲ ۳۳ ۳۴ ۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰ ۴۱ ۴۲ ۴۳ ۴۴ ۴۵ ۴۶ ۴۷ ۴۸ ۴۹ ۵۰ ۵۱
 Ankl.14 | ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ ۲۸ ۲۹ ۳۰ ۳۱ ۳۲ ۳۳ ۳۴ ۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰ ۴۱ ۴۲ ۴۳ ۴۴ ۴۵ ۴۶ ۴۷ ۴۸ ۴۹ ۵۰ ۵۱
 ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ ۲۸ ۲۹ ۳۰ ۳۱ ۳۲ ۳۳ ۳۴ ۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰ ۴۱ ۴۲ ۴۳ ۴۴ ۴۵ ۴۶ ۴۷ ۴۸ ۴۹ ۵۰ ۵۱
 ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ ۲۸ ۲۹ ۳۰ ۳۱ ۳۲ ۳۳ ۳۴ ۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰ ۴۱ ۴۲ ۴۳ ۴۴ ۴۵ ۴۶ ۴۷ ۴۸ ۴۹ ۵۰ ۵۱
 ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ ۲۸ ۲۹ ۳۰ ۳۱ ۳۲ ۳۳ ۳۴ ۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰ ۴۱ ۴۲ ۴۳ ۴۴ ۴۵ ۴۶ ۴۷ ۴۸ ۴۹ ۵۰ ۵۱

Figure 4.22: A page of typeset Pahlavi text from *Zand of Bahman Yasn* [18]. Note the upside-down *Ahremen* (𐭡𐭥𐭫𐭮)

زنده‌یمن	۹۰
- واژه‌های ترکیبی در ردیف الفبایی خود آمده‌اند و در زیر اجزاء خود با نشانه ← بصورت ترکیبی بازگشت داده شده‌اند.	
- حرفهای اضافه و ربط مرکب بنا بر ترتیب الفبائی خود آمده‌اند مثلاً	
۶۵ ۶۶ ۶۷ در زیر ۶۶ و ۱۲۳۴ ۱۲۳۵ در زیر ۱۲۳۳ آمده است	
«...»	
۱۲۳۴	xānīg «خانی، چشمه» ۱، ۳۶
۳۱۲۳۴	xānīgān (ج) «چشمه‌ها» ۹، ۳۰
۱۲۳	any «دیگر» ۷، ۳
۱۲۳/۱	āhōg «آهو، عیب» ۱، ۲۹؛ ۶، ۲۹
۱۲۳۴	a-hōš «بی مرگ» ۶، ۵؛ ۶، ۵؛ ۷، ۲؛ ۷، ۲؛ ۶، ۷
۱۲۳۴۵	a-hōšt «بی مرگی» ۲، ۳؛ ۳، ۶
۱/۱۲۳۴۵	hāwišt «شاگرد» ۶۲، ۶؛ ۷۲؛ ۲
۱/۳۱۲۳۴۵	hāwištān (ج) «شاگردان» ۲۹، ۵؛ ۳۷؛ ۸
۱۲۳۴۵	hāwištīh «شاگردی» ۲۸، ۸
۱۲۳۴	pas «پس» ۲، ۳؛ ۳، ۴۵؛ ۴، ۵۱؛ ۷، ۶۳؛ ۱، ۷۷؛ ۶
۱۲۳۴۵۶	ahlāyīh «پرهیزگاری، پارسایی» ۴۳، ۷؛ ۴۳، ۹
۱۲۳۴۵	ahlaw «پرهیزگار، پارسا» ۸، ۳؛ ۵؛ ۷، ۳۹؛ ۳، ۴۴؛ ۵؛ ۵۲؛ ۸ ←
۱۲۳۴۵۶	ahlawān (ج) «پرهیزگاران» ۴۱، ۴؛ ۲۸، ۲
۱۲۳۴۵۶	ahlawtar «پرهیزگارتر» ۶، ۵
۱۲۳۴۵۶	ahlaw-dād «اهل‌داد، صدقه» ۳۳، ۱
۱۲۳۴۵۶	ahreman «اهریمن، نیروی ویرانگری که در برابر اورمزد و آفریده‌های او ایستاده و نیرو آراسته است» ۱۳، ۱؛ ۱۵؛ ۴، ۲۹؛ ۲
۱۲۳۴۵	ahlamōg «اهلموغ، مرتد، بدعت‌گذار» ۷۷، ۶؛ ۷۷، ۸؛ ۷۸، ۹

Figure 4.23: A page from the glossary of [18]. Note the word *Ahreman* written regularly as ۱۲۳۴۵۶

		PAHLAVI KEY		
	[4]		[3]	
xām		𐬰𐬀		𐬀𐬀𐬀𐬀𐬀
hāmharz		𐬀𐬀𐬀𐬀𐬀		𐬀𐬀𐬀
hāmīnīg		𐬀𐬀𐬀𐬀		𐬀𐬀𐬀𐬀
hāmīn		𐬀𐬀𐬀		𐬀𐬀𐬀𐬀
xāmīz		𐬀𐬀𐬀		𐬀𐬀𐬀𐬀
hāmwār		𐬀𐬀𐬀𐬀		𐬀𐬀
hāmōyēn		𐬀𐬀𐬀𐬀		𐬀𐬀
hāmōn		𐬀𐬀𐬀		𐬀𐬀
xāmōš		𐬀𐬀𐬀		𐬀𐬀𐬀
hāmkišwar		𐬀𐬀𐬀𐬀𐬀		𐬀𐬀𐬀𐬀
hāmist		𐬀𐬀𐬀𐬀		𐬀𐬀𐬀𐬀
hās ² r		𐬀𐬀𐬀		𐬀𐬀𐬀
āxīstan		𐬀𐬀𐬀𐬀		𐬀𐬀𐬀𐬀𐬀
hāz-		-𐬀𐬀		𐬀𐬀
hāzišn		𐬀𐬀𐬀𐬀		𐬀𐬀𐬀
² hād, haxt		𐬀𐬀𐬀		𐬀𐬀𐬀𐬀𐬀
hādōxt		𐬀𐬀𐬀𐬀𐬀		Ahreman 𐬀𐬀𐬀 / 𐬀𐬀𐬀
hādamāns ² r		𐬀𐬀𐬀𐬀𐬀𐬀		Ahrišwang 𐬀𐬀𐬀𐬀
hādamāns ² rīg		𐬀𐬀𐬀𐬀𐬀𐬀𐬀		ahlawīh 𐬀𐬀𐬀
axtar		𐬀𐬀𐬀		ahlawdād 𐬀𐬀𐬀𐬀
axtarāmār		𐬀𐬀𐬀𐬀𐬀		ahlaw 𐬀𐬀
axtarmār		𐬀𐬀𐬀𐬀		Ahreman 𐬀𐬀 / 𐬀𐬀
axtarmārīh		𐬀𐬀𐬀𐬀𐬀		ahlomōyīh 𐬀𐬀𐬀
xwah ¹		𐬀𐬀𐬀		ahlomōy 𐬀𐬀
xwahar ¹		𐬀𐬀𐬀𐬀		xārpušt 𐬀𐬀𐬀𐬀
				āhanjīdan 𐬀𐬀𐬀𐬀
				ahōš 𐬀𐬀
				ahōšīh 𐬀𐬀𐬀
				hāwištīh 𐬀𐬀𐬀𐬀
				hāwišt 𐬀𐬀𐬀
				xāk 𐬀𐬀
				āhr, xār 𐬀𐬀
				pas ¹ 𐬀𐬀
				ahlā 𐬀𐬀
				ahlāyīh 𐬀𐬀𐬀
				xārōmand 𐬀𐬀𐬀
				ahrām-, āxrām- 𐬀𐬀
				āxrāmīdan 𐬀𐬀𐬀𐬀
				ahrāftan 𐬀𐬀𐬀
				pasīh ¹ 𐬀𐬀
				pasdānišnīh ¹ 𐬀𐬀𐬀𐬀

Figure 4.24: A page from [13]. Note the two different spellings for *Ahreman* (red). Each spelling variant has an upside-down form as well.

Bibliography

- [1] M Abol-Ghassemi. *A Manual of Old Iranian Languages (Part I)*. Samt, 1996.
- [2] J Amoozgar and A Tafazzoli. *Pahlavi Language, Literature, Grammatical Sketch, Texts and Glossary*. Moin, 1996.
- [3] Ervad Tahmuras Dinshaji Anklesaria, editor. *The Bûndahishn*. British India Press, 1908.
- [4] rahām Aša, editor. *Hormazd o HarvīspĀgāhī*. Asātīr, Tehran, 1390 (H.S.).
- [5] The Unicode Consortium. *The Unicode Standard Version 6.3.0*. the Unicode Consortium, Mountain View, CA, 2013.
- [6] Ervad Bamanji Nasarvanji Dhabhar, editor. *The Epistles of Mânûshchîhar*. Trustees of the Parsee Panchayat Funds and Properties, Bombay, 1912.
- [7] Michael Everson, Roozbeh Pournader, and Desmond Durkin-Meisterenst. Preliminary proposal to encode the Book Pahlavi script in the BMP of the UCS. *ISO/IEC JTC1/SC2/WG2 N3294 and UTC Document Register L2/07-234*, 2007.
- [8] Hoshang Jamasp and Gandevia Mervanji Manekji, editors. *Vendidâd*. Government of Bombay, Bombay, 1907.
- [9] Jâmâsp-Âsânâ and West, editors. *Shikand-Gûmânîk Vijâr*. Government of Bombay, 1887.
- [10] Jamaspji Dastur Minocherji Jamasp-Asana, editor. *Corpus of Pahlavi Texts*. -, 1913.

- [11] Hoshengji Jamaspji and Martin Haug, editors. *An Old Zand-Pahlavi Glossary*. Government of Bombay, Bombay, 1867.
- [12] Antiâ Ervad Edalji Kersâspji, editor. *Pâzend Texts*. The trustees of the Parsee Punchâyet, Bombay, 1909.
- [13] D. N. MacKenzie. *A Concise Pahlavi Dictionary*. Oxford University Press, 1986.
- [14] Dhanjishah Meherjibhai Madan, editor. *The complete text of the Pahlavi Dinkard*. The society for the promotion of researches into the Zoroastrian religion, 1911.
- [15] Katâyûn Mazdâpûr, editor. *Dâstân-e Garšâsp, Tahmûres o Jamšîd, Gelšâh o Matnhâ-ye dîgar*. Âgâh, Tehran, 1378 (H.S).
- [16] Mahshîd Mîrfakhrâi, editor. *Hâdôxt Nask*. Institute of Humanities and Cultural Studies, Tehrân, 2007.
- [17] M Mirfakhraie, editor. *Baγân Yasn: Avestan and Zand*. Institute for Humanities and Cultural Studies, Tehran, 2003.
- [18] Mohammad Taqî Râshed Mohassel, editor. *Zand-e Bahman Yasn: Edited, Transcribed, Translated and Annotated*. Institute for Humanities and Cultural Studies, Tehran, 2006.
- [19] H. S. Nyberg. *A Manual of Pahlavi*. Otto Harrassowitz, 1964, 1974.
- [20] Roozbeh Pournader. Preliminary proposal to encode the Book Pahlavi script in the unicode standard. *UTC Document Register L2/13-141*, 2013.
- [21] M. T. Rashed Mohassel, editor. *Srôš Yasn: Avestan and Zand of Yasn 57*. Institute for Humanities and Cultural Studies, Tehrân, 2002.
- [22] P. O. Skjaervo. *Introduction to Pahlavi*. Cambridge, Mass., 2007.
- [23] Peshoutun Dustoor Behramjee Sunjana, editor. *The Dinkard*. Duftur Ashkara, Bombay, 1883.
- [24] Unknown. Dâdistan-ê Dînik; Rivâyat ; Mânûstshîhar ; Zâdsparam. <http://www.kb.dk/manus/ortsam/2009/okt/orientalia/object68131/en/>, 1572.

- [25] Unknown. Minoqäräd; sarosh-yast. <http://www.kb.dk/manus/ortsam/2009/okt/orientalia/da/object64031/>, 1700–1799.
- [26] Unknown. Rivayat etc. <http://www.kb.dk/manus/ortsam/2009/okt/orientalia/da/object65997/>, 1700–1799.
- [27] Unknown. Sifat-i-sirozah, Pahlavicum, Izeshne rafitwan, Afrigan rapithwan, Afrin i Zartust paigambar, Sirozah e qorda avesta. <http://www.kb.dk/manus/ortsam/2009/okt/orientalia/da/object65491/>, 1700/1799.
- [28] Unknown. Viraf-namah, Bundähäsh etc. <http://www.kb.dk/manus/ortsam/2009/okt/orientalia/object63895/en/>, 1700/1799.
- [29] Unknown. Sifat-i-sirozah. <http://www.kb.dk/manus/ortsam/2009/okt/orientalia/da/object65189/>, 1800/1820.
- [30] UTC. Unicode FAQ: Ligatures, digraphs, presentation forms vs plain text. http://www.unicode.org/faq/ligature_digraph.html. Accessed: 2014-01-01.
- [31] E. W. West and Martin Haug. *Glossary and Index of Pahlavi texts*. Government of Bombay, 1874.

**ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹**

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html> for guidelines and details before filling this form.

Please ensure you are using the latest Form from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest Roadmaps.

A. Administrative

1. Title:	<i>Proposal for Encoding Book Pahlavi in the Unicode Standard</i>		
2. Requester's name:	<i>Abe Meyers</i>		
3. Requester type (Member body/Liaison/Individual contribution):	<i>Individual Contribution</i>		
4. Submission date:	<i>March 3, 2014</i>		
5. Requester's reference (if applicable):	<i>N/A</i>		
6. Choose one of the following:			
This is a complete proposal:			<i>Yes</i>
(or) More information will be provided later:			<i>No</i>

B. Technical – General

1. Choose one of the following:			
a. This proposal is for a new script (set of characters):			<i>Yes</i>
Proposed name of script:	<i>Book Pahlavi</i>		
b. The proposal is for addition of character(s) to an existing block:			<i>Yes</i>
Name of the existing block:	<i>Avestan</i>		
2. Number of characters in proposal:			<i>32 (30 + 2)</i>
3. Proposed category (select one from below - see section 2.2 of P&P document):			
A-Contemporary	<input type="checkbox"/> B.1-Specialized (small collection)	<input type="checkbox"/> B.2-Specialized (large collection)	
C-Major extinct	<input checked="" type="checkbox"/> D-Attested extinct	<input type="checkbox"/> E-Minor extinct	
F-Archaic Hieroglyphic or Ideographic		G-Obscure or questionable usage symbols	
4. Is a repertoire including character names provided?			<i>Yes</i>
a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?			<i>Yes</i>
b. Are the character shapes attached in a legible form suitable for review?			<i>Yes</i>
5. Fonts related:			
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	<i>Abraham Meyers</i>		
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	<i>Abraham Meyers</i>		
6. References:			
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?			<i>Yes</i>
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?			<i>Yes</i>
7. Special encoding issues:			
Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?			<i>Yes</i>

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at <http://www.unicode.org> for such information on other scripts. Also see Unicode Character Database (<http://www.unicode.org/reports/tr44/>) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

¹ Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	Yes
If YES explain	<i>They were preliminary proposals</i>
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?	
If YES, with whom?	
If YES, available relevant documents:	
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	
Reference:	
4. The context of use for the proposed characters (type of use; common or rare)	Common
Reference:	
5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference:	<i>Scholarly community, education, The Zoroastrian community</i>
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	No
If YES, is a rationale provided?	
If YES, reference:	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	Yes
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	No
If YES, is a rationale for its inclusion provided?	N/A
If YES, reference:	N/A
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	No
If YES, is a rationale for its inclusion provided?	N/A
If YES, reference:	
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?	Yes
If YES, is a rationale for its inclusion provided?	<i>They have different identities</i>
If YES, reference:	
11. Does the proposal include use of combining characters and/or use of composite sequences?	Yes
If YES, is a rationale for such use provided?	
If YES, reference:	<i>Yes. See proposal</i>
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	N/A
If YES, reference:	N/A
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	No
If YES, describe in detail (include attachment if necessary)	N/A
13. Does the proposal contain any Ideographic compatibility characters?	No
If YES, are the equivalent corresponding unified ideographic characters identified?	N/A
If YES, reference:	N/A