Jost Gippert

The Encoding of Avestan – Problems and Solutions

Abstract

‘Avestan’ is the name of the ritual language of Zoroastrianism, which was the state religion of the Iranian empire in Achaemenid, Arsacid and Sasanid times, covering a time span of more than 1200 years.[1] It is named after the ‘Avesta’, i.e., the collection of holy scriptures that form the basis of the religion which was allegedly founded by Zarathushtra, also known as Zoroaster, by about the beginning of the first millennium B.C. Together with Vedic Sanskrit, Avestan represents one of the most archaic witnesses of the Indo-Iranian branch of the Indo-European languages, which makes it especially interesting for historical-comparative linguistics. This is why the texts of the Avesta were among the first objects of electronic corpus building that were undertaken in the framework of Indo-European studies, leading to the establishment of the TITUS database (‘Thesaurus indogermanischer Text- und Sprachmaterialien’).[2] Today, the complete Avestan corpus is available, together with elaborate search functions[3] and an extended version of the subcorpus of the so-called ‘Yasna’, which covers a great deal of the attestation of variant readings.[4]

Right from the beginning of their computational work concerning the Avesta, the compilers[5] had to cope with the fact that the texts contained in it have been transmitted in a special script written from right to left, which was also used for printing them in the scholarly editions used until today.[6] It goes without saying that there was no way in the middle of the 1980s to encode the Avestan scriptures exactly as they are found in the manuscripts. Instead, we had to rely upon transcriptional devices that were dictated by the restrictions of character encoding as provided by the computer systems used. As the problems we had to face in this respect and the solutions we could apply are typical for the development of computational work on ancient languages, it seems worthwhile to sketch them out here.

1 The Avestan script and its transcription

1.1 Early western approaches to the Avestan script and its transcription

The Avestan script has been known to western scholarship since the 17th century when the first accounts of the religion of the ‘Parsees’, i.e., Zoroastrians living in India and Iran, were published. The first notable description of the script is found in the travel report by JEAN CHARDIN who sojourned in Iran in 1673–7; in the 1711 edition of his report,[7] the author provides an ‘alphabet of the ancient Persians’, together with a lithographed table contrasting the characters of the Avestan script with their Perso-Arabian equivalents;[8] cf. the extract illustrated in Fig. 1.[9]

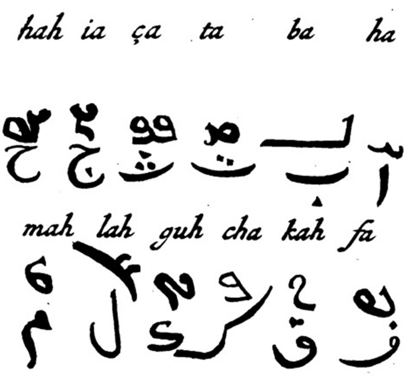
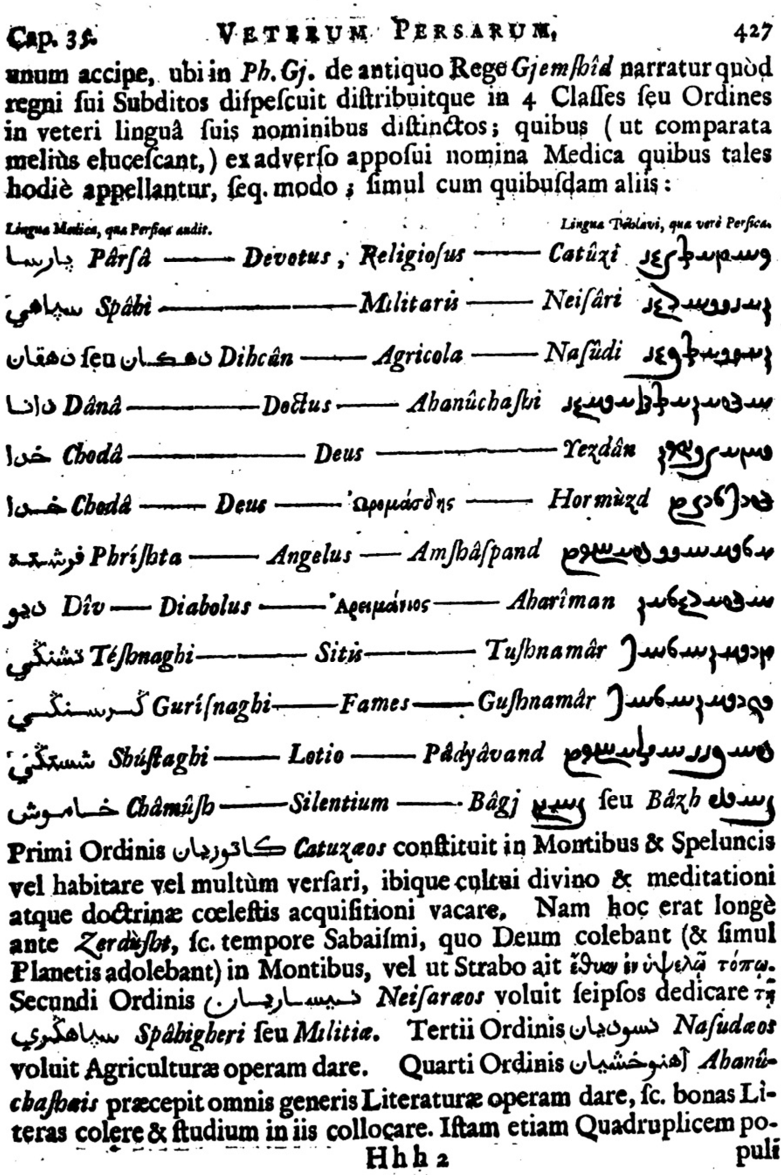


Fig. 1: CHARDIN's alphabet list (extract)

A much more interesting account than CHARDIN's,[10] who took the Persian letters to be 'small' variants of the 'big' Avestan ones,[11] is that of THOMAS HYDE who in his 'History of the Religion of the Ancient Persians, Parthians and Medes' of 1700 provided the first specimens of words written in Avestan characters along with a Latin transcription. The words are not in Avestan (or 'Zend'), however, but 'in the Pahlavi language, which is verily Persian' (cf. Fig. 2);[12] as a matter of fact, what we have here is a list of words in 'Pazend', i.e. Middle Persian (Pahlavi) written in Avestan characters.[13]

To the (postumous) second edition (of 1760) of HYDE's work, the editor (G. COSTARD) added a comprehensive Table displaying the 'Letters used in the books in Zend and Pazend, according to the copies by DR. HYDE, together with the Zend ligatures and abbreviations' *in toto*, together with a detailed explanation of their values in Latin script (cf. Fig. 3).[14]

An even more detailed account of 'Zend' and 'Pehlvi' characters was published by (ABRAHAM-HYACINTHE) ANQUETIL-DUPERRON in his comprehensive treatise on the 'Zend-Avesta' in 1771 (cf. Fig. 4).[15] ANQUETIL's description, which was derived from two manuscripts of the Bibliothèque Nationale in Paris, is generally regarded as the beginning of modern Avestology. The transcription he used was clearly based upon French orthography. In a similar way, IGNACY PIETRASZEWSKI in 1858 explicitly applied Polish rules to his transcripts (cf. Fig. 5 and Fig. 6).[16]



Cap. 35. VETERUM PERSARUM, 427

unum accipe, ubi in *Ph. Gj.* de antiquo Rege *Gjemſhîd* narratur quòd regni ſui Subditos diſpeſcuit diſtribuitque in 4 Claſſes ſeu Ordines in veteri linguâ ſuis nominibus diſtinctos; quibus (ut comparata meliùs elluceſcant,) ex adverſo appoſui nomina Medica quibus tales hodiè appellantur, ſeq. modo; ſimul cum quibuſdam aliis:

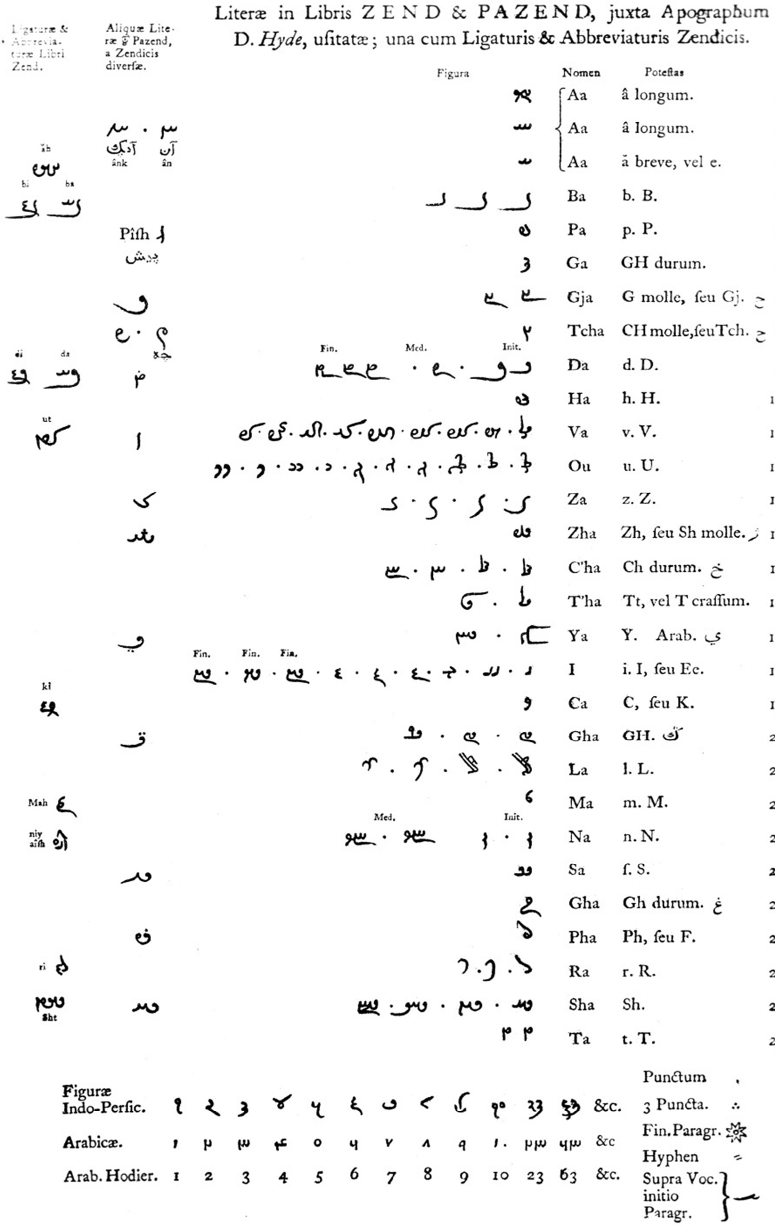
Lingua Medica, quæ Perſica audit. *Lingua Pehlavi, quæ verè Perſica.*

پارسا	*Pârſâ*	*Devotus, Religioſus*		*Catûzî*	
سپاهي	*Spâhi*	*Militaris*		*Neiſâri*	
دهقان ſeu دهكان	*Dihcân*	*Agricola*		*Naſûdi*	
دانا	*Dânâ*	*Doctus*		*Abanûchaſhi*	
خدا	*Chodâ*	*Deus*		*Yezdân*	
خدا	*Chodâ*	*Deus*	Ὡρομάσδης	*Hormùzd*	
فرشته	*Phriſhta*	*Angelus*		*Amſhâſpand*	
ديو	*Dîv*	*Diabolus*	Ἀρειμάνιος	*Aharîman*	
تشنگي	*Téſhnaghi*	*Sitis*		*Tuſhnamâr*	
گرسنگي	*Guriſnaghi*	*Fames*		*Guſhnamâr*	
شستگي	*Shúſtaghi*	*Lotio*		*Pâdyâvand*	
خاموش	*Châmûſh*	*Silentium*		*Bâgj* seu *Bâzh*	

Primi Ordinis كاتوزيان *Catuzæos* conſtituit in Montibus & Speluncis vel habitare vel multùm verſari, ibique cultui divino & meditationi atque doctrinæ cœleſtis acquiſitioni vacare. Nam hoc erat longè ante *Zerdùſht*, ſc. tempore Sabaiſmi, quo Deum colebant (& ſimul Planetis adolebant) in Montibus, vel ut Strabo ait ἔθυον ἐν ὑψηλῷ τόπῳ. Secundi Ordinis نيسارييان *Neiſaræos* voluit ſeipſos dedicare τῇ سپاهگري *Spâhigheri* ſeu *Militiæ*. Tertii Ordinis نسوديان *Naſudæos* voluit Agriculturæ operam dare. Quarti Ordinis آهنوخشيان *Abanûchaſhæis* præcepit omnis generis Literaturæ operam dare, ſc. bonas Literas colere & ſtudium in iis collocare. Iſtam etiam Quadruplicem populi

Hhh 2

Fig. 2: Pazend word list in HYDE, Historia, 1711



Literæ in Libris ZEND & PAZEND, juxta Apographum D. *Hyde*, uſitatæ; una cum Ligaturis & Abbreviaturis Zendicis.

Ligaturæ & Abbreviaturæ Libri Zend.
Aliquæ Literæ & Pazend, a Zendicis diverſæ.
ăh
آدنگ آن
ânk ân
bi ba
Pîſh
di da
ut
kl
Mah
niy aiſh
ri
Sht
Fin. Med. Init.
Fin. Fin. Fin.

Figura	Nomen	Poteſtas	
	Aa	â longum.	1
	Aa	â longum.	2
	Aa	ă breve, vel e.	3
	Ba	b. B.	4
	Pa	p. P.	5
	Ga	GH durum.	6
	Gja	G molle, ſeu Gj. ج	7
	Tcha	CH molle,ſeuTch. چ	8
	Da	d. D.	9
	Ha	h. H.	10
	Va	v. V.	11
	Ou	u. U.	12
	Za	z. Z.	13
	Zha	Zh, ſeu Sh molle. ژ	14
	C'ha	Ch durum. خ	15
	T'ha	Tt, vel T craſſum.	16
	Ya	Y. Arab. ي	17
	I	i. I, ſeu Ee.	18
	Ca	C, ſeu K.	19
	Gha	GH. گ	20
	La	l. L.	21
	Ma	m. M.	22
	Na	n. N.	23
	Sa	ſ. S.	24
	Gha	Gh durum. غ	25
	Pha	Ph, ſeu F.	26
	Ra	r. R.	27
	Sha	Sh.	28
	Ta	t. T.	29

Figuræ Indo-Perſic.	૧	૨	૩	૪	૫	૬	૭	૮	૯	૧૦	૨૩	૬૩	&c.
Arabicæ.	١	٢	٣	۴	٥	٦	٧	٨	٩	١٠	٢٣	٦٣	&c
Arab. Hodier.	1	2	3	4	5	6	7	8	9	10	23	63	&c.

Punctum	.
3 Puncta.	∴
Fin.Paragr.	✲
Hyphen	=
Supra Voc. initio Paragr.	

Fig. 3: Alphabet list in HYDE, Historia, 1760.

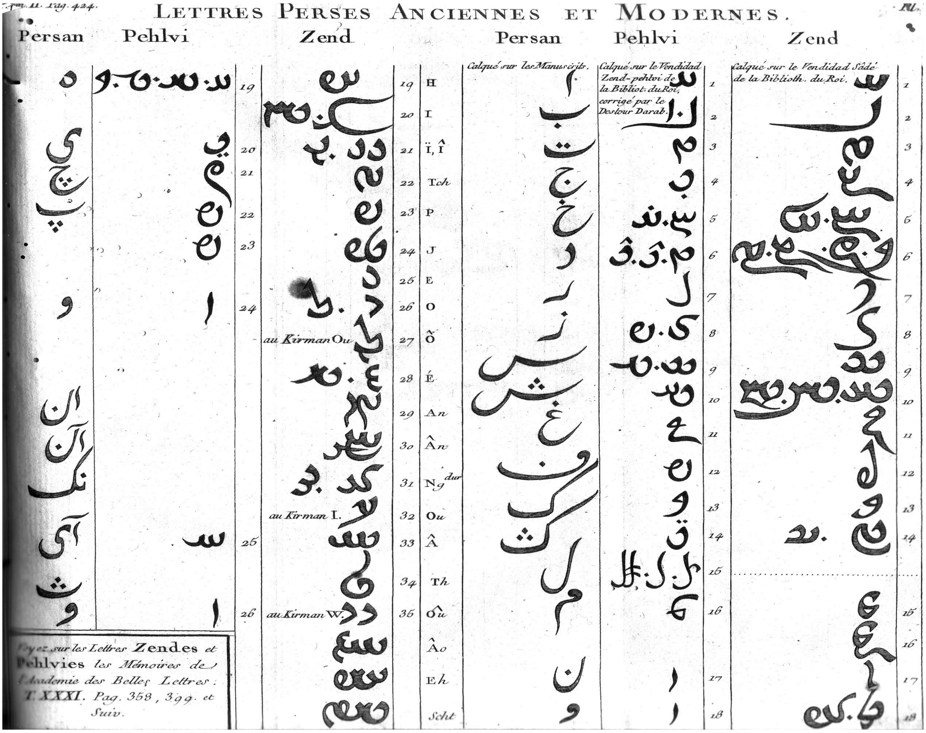
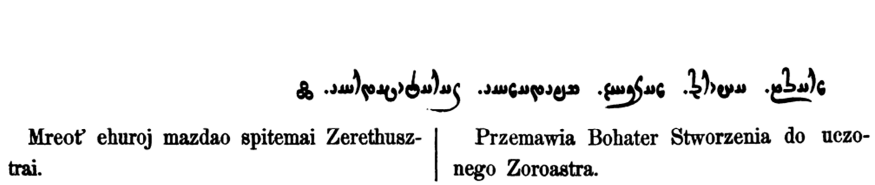
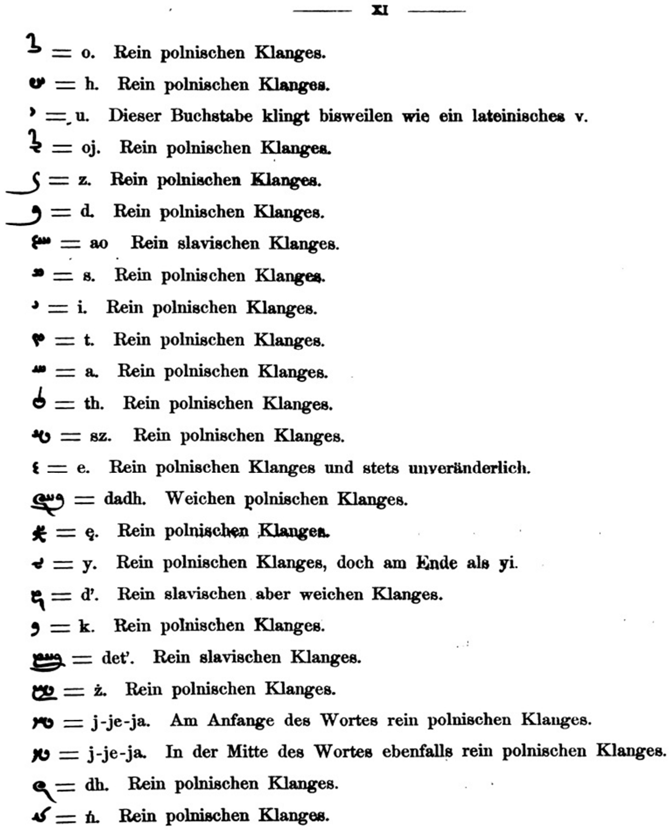

LETTRES PERSES ANCIENNES ET MODERNES.
Persan
Pehlvi
Zend
Persan
Pehlvi
Zend
Calqué sur les Manuscrits.
Calqué sur le Vendidad Zend-pehlvi de la Biblioth. du Roi, corrigé par le Destour Darab.
Calqué sur le Vendidad Sadé de la Biblioth. du Roi.
H
I
Ï, Î
Tch
P
J
E
O
Ô
É
An
Ân
Ng dur
Ou
Â
Th
Ôu
Âo
Eh
Scht
au Kirman Ou
au Kirman I.
au Kirman W.
Voyez sur les Lettres Zendes et Pehlvies les Mémoires de l'Academie des Belles Lettres: T. XXXI. Pag. 358, 399. et Suiv.

Fig. 4: Alphabet list in ANQUETIL-DUPERRON, Zend-Avesta



Mreot' ehuroj mazdao spitemai Zerethusztrai. | Przemawia Bohater Stworzenia do uczonego Zoroastra.

Fig. 5: PIETRASZEWSKI'S transcription (and Polish translation) of Vd. 1,1.

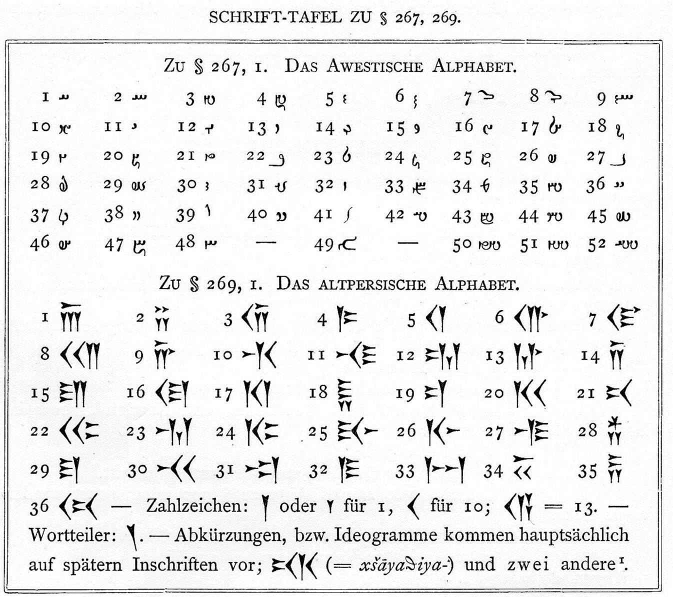


— XI —

= o. Rein polnischen Klanges.
= h. Rein polnischen Klanges.
= u. Dieser Buchstabe klingt bisweilen wie ein lateinisches v.
= oj. Rein polnischen Klanges.
= z. Rein polnischen Klanges.
= d. Rein polnischen Klanges.
= ao Rein slavischen Klanges.
= s. Rein polnischen Klanges.
= i. Rein polnischen Klanges.
= t. Rein polnischen Klanges.
= a. Rein polnischen Klanges.
= th. Rein polnischen Klanges.
= sz. Rein polnischen Klanges.
= e. Rein polnischen Klanges und stets unveränderlich.
= dadh. Weichen polnischen Klanges.
= ę. Rein polnischen Klanges.
= y. Rein polnischen Klanges, doch am Ende als yi.
= d'. Rein slavischen aber weichen Klanges.
= k. Rein polnischen Klanges.
= det'. Rein slavischen Klanges.
= ż. Rein polnischen Klanges.
= j-je-ja. Am Anfange des Wortes rein polnischen Klanges.
= j-je-ja. In der Mitte des Wortes ebenfalls rein polnischen Klanges.
= dh. Rein polnischen Klanges.
= ń. Rein polnischen Klanges.

Fig. 6: PIETRASZEWSKI's transcription

The different approaches to a Romanization of the Avestan script had reached a preliminary end by the beginning of the 20th century when CHRISTIAN BARTHOLOMAE, first in his account on the Avestan and Old Persian languages in W. GEIGER's and E. KUHN's 'Outline of Iranian Philology' (1895-1901) and then in his 'Old Iranian Dictionary' (1904), proposed a transcription system that was based upon the choice of original characters used in K. F. GELDNER's edition of the Avesta (cf. Fig. 7 – Fig. 9). Due to the importance of the dictionary, which has remained the standard reference work of Avestan lexicography until the present day, BARTHOLOMAE's transcription system was used for many years to come.



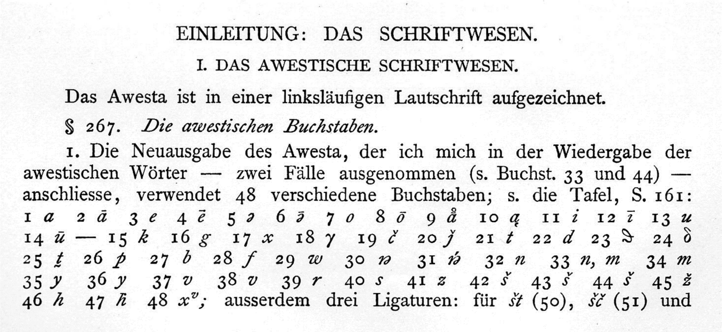
SCHRIFT-TAFEL ZU § 267, 269.

ZU § 267, 1. DAS AWESTISCHE ALPHABET.

ZU § 269, 1. DAS ALTPERSISCHE ALPHABET.

36 — Zahlzeichen: 𐏑 oder 𐏑 für 1, 𐏓 für 10; 𐏓𐏑𐏒 = 13. — Wortteiler: 𐏐. — Abkürzungen, bzw. Ideogramme kommen hauptsächlich auf spätern Inschriften vor; 𐏈 (= *xšāyaϑiya-*) und zwei andere[1].

Fig. 7: The Avestan alphabet in BARTHOLOMAE (1895-1901: 161)



EINLEITUNG: DAS SCHRIFTWESEN.

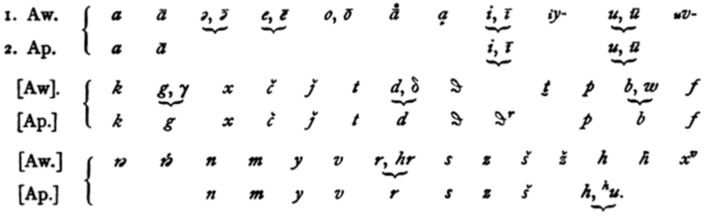
I. DAS AWESTISCHE SCHRIFTWESEN.

Das Awesta ist in einer linksläufigen Lautschrift aufgezeichnet.

§ 267. *Die awestischen Buchstaben.*

1. Die Neuausgabe des Awesta, der ich mich in der Wiedergabe der awestischen Wörter — zwei Fälle ausgenommen (s. Buchst. 33 und 44) — anschliesse, verwendet 48 verschiedene Buchstaben; s. die Tafel, S. 161: 1 *a* 2 *ā* 3 *e* 4 *ē* 5 *ə* 6 *ə̄* 7 *o* 8 *ō* 9 *å̄* 10 *ą* 11 *i* 12 *ī* 13 *u* 14 *ū* — 15 *k* 16 *g* 17 *x* 18 *γ* 19 *č* 20 *ǰ* 21 *t* 22 *d* 23 *ϑ* 24 *δ* 25 *t̰* 26 *p* 27 *b* 28 *f* 29 *w* 30 *ŋ* 31 *ŋ́* 32 *n* 33 *n, m* 34 *m* 35 *y* 36 *y* 37 *v* 38 *v* 39 *r* 40 *s* 41 *z* 42 *š* 43 *š* 44 *š* 45 *ž* 46 *h* 47 *h̄* 48 *xᵛ*; ausserdem drei Ligaturen: für *št* (50), *šč* (51) und

Fig. 8: Transcription according to BARTHOLOMAE (1895-1901: 152)



1. Aw.	*a*	*ā*	*ə, ə̄*	*e, ē*	*o, ō*	*å̄*	*ą*	*i, ī*	*iy-*	*u, ū*	*uv-*
2. Ap.	*a*	*ā*						*i, ī*		*u, ū*	

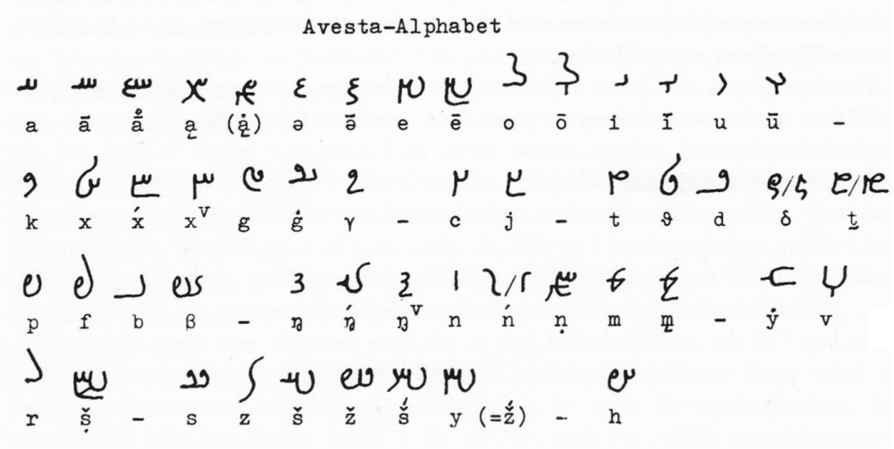
[Aw].	*k*	*g, γ*	*x*	*č*	*ǰ*	*t*	*d, δ*	*ϑ*		*t̰*	*p*	*b, w*	*f*
[Ap.]	*k*	*g*	*x*	*č*	*ǰ*	*t*	*d*	*ϑ*	*ϑʳ*		*p*	*b*	*f*

[Aw.]	*ŋ*	*ŋ́*	*n*	*m*	*y*	*v*	*r, hr*	*s*	*z*	*š*	*ž*	*h*	*h̄*	*xᵛ*
[Ap.]			*n*	*m*	*y*	*v*	*r*	*s*	*z*	*š*		*h, ʰu.*		

Fig. 9: Transcriptional alphabet in BARTHOLOMAE (1904: xxiii)

1.2 The ‘Hoffmann system’

On the basis of a thorough reconsideration of the character inventory and its linguistic background, BARTHOLOMAE’s system was challenged to a certain extent by KARL HOFFMANN (cf. Fig. 10).[17] It is HOFFMANN’s merit to have clarified the function and mutual relationship of the three characters numbered 42–44, all transcribed by plain š in BARTHOLOMAE’s works, as well as several other letters. Table 1 illustrates the peculiarities of the system thus achieved, which was the first to be strictly transliterative in the sense that all characters (rather: graphemes) of the original script are rendered by one unique Latin symbol, in contrast to the ‘mixed’ systems of former authors.[18]



Avesta-Alphabet

a ā ā̊ ą (ą̇) ə ə̄ e ē o ō i ī u ū –

k x x́ xᵛ g ǵ γ – c j – t ϑ d δ t̰

p f b β – ŋ ŋ́ ŋᵛ n ń ṇ m m̨ – ẏ v

r ṣ̌ – s z š ž š́ y (=ž́) – h

Fig. 10: HOFFMANN’s ‘Zeicheninventar’

Original	Anqu.-No.	Anqu.	Pietr.	Bthl.-No.	Bthl.	Hoffmann	Original
𐬀	1	a, e	e	1	*a*	*a*	𐬀
𐬁	33	â	a	2	*ā*	*ā*	𐬁
𐬂	—	—	—	—	—	*å*	𐬂
𐬃	(36)	âo	ao	9	*å*	*ā̊*	𐬃
𐬄	29	an	—	10	*ą*	*ą*	𐬄
𐬅	—	—	ę	—	—	*ą̇*	𐬅
𐬅 [19]	—	—	—	—	—		𐬅 [19]
𐬆	28	e	e	5	*ə*	*ə*	𐬆
𐬇	—	—	—	6	*ə̄*	*ə̄*	𐬇
𐬈	28	e	j-je-ja	3	*e*	*e*	𐬈
𐬉	—	—	je	4	*ē*	*ē*	𐬉
𐬊	26	o	o	7	*o*	*o*	𐬊

Original	Anqu.-No.	Anqu.	Pietr.	Bthl.-No.	Bthl.	Hoffmann	Original
𐬋	27	ô	oj	8	*ō*	*ō*	𐬋
𐬌	25	e	i	11	*i*	*i*	𐬌
𐬍	21	ï, î	y	12	*ī*	*ī*	𐬍
𐬎	26	o	u	13	*u*	*u*	𐬎
𐬏	32	ou	uj	14	*ū*	*ū*	𐬏
𐬐	13	k, c	k	15	*k*	*k*	𐬐
𐬑	5	kh	ch	17	*x*	*x*	𐬑
𐬒				47	*ḣ*	*x́*	𐬒
𐬓	—	—	—	48	x^v	x^v	𐬓
𐬔	14	g^{dur}	g	16	*g*	*g*	𐬔
𐬕			—	—	—	*ġ*	𐬕
𐬖	11	gh	gh	18	*γ*	*γ*	𐬖
𐬗	22	tch	cz	19	*č*	*c*	𐬗
𐬘	4	dj	dz, dż	20	*ǰ*	*j*	𐬘
𐬙	3	t	t	21	*t*	*t*	𐬙
𐬚	34	th	th	23	*ϑ*	*ϑ*	𐬚
𐬛	6	d	d	22	*d*	*d*	𐬛
𐬜			d’	24	*δ*	*δ*	𐬜
𐬜[19]			dh	—	—		𐬜[19]
𐬝	—	—	t’	25	*t̰*	*t̰*	𐬝
𐬞	23	p	p	26	*p*	*p*	𐬞
𐬟	12	f	f	28	*f*	*f*	𐬟
𐬠	2	b	b	27	*b*	*b*	𐬠
𐬡	18	v	w	29	*w*	*β*	𐬡
𐬢	31	ng^{dur}	ṅ	30	*ŋ*	*ŋ*	𐬢
𐬣			ń	31	*ŋ́*	*ŋ́*	𐬣
𐬤	—	—	—	—	—	$ŋ^v$	𐬤
𐬥	17	n	n	32	*n*	*n*	𐬥
𐬦	—	—	—	—	—	*ń*	𐬦
𐬦[19]	—	—	—	—	—		𐬦[19]
𐬧	30	ân	ą	33	*n,m*	*ṇ*	𐬧
𐬨	15	m	m	34	*m*	*m*	𐬨
𐬩	16	hm	ehm	—	—	*m̨*	𐬩

Original	Anqu.- No.	Anqu.	Pietr.	Bthl.- No.	Bthl.	Hoffmann	Original
𐬪	20	i	—	49	*y*	*ẏ*	𐬪
𐬫			j-je-ja	35		*y*	𐬫
𐬌𐬌	21	ï, î	y	36		*ii*	𐬌𐬌[20]
𐬬	18	v	—	37	*v*	*v*	𐬬
[19]	—	—	—	—			[19]
𐬎𐬎	35	o^u	w	38		*uu*	𐬎𐬎[20]
𐬭	7	r	r	39	*r*	*r*	𐬭
𐬮	—	(l)	(l)	—	—	—	𐬮
𐬯	9	s	s	40	*s*	*s*	𐬯
𐬰	8	z	z	41	*z*	*z*	𐬰
𐬱	10	sch	sz	42	*š*	*š*	𐬱
𐬳			—	44		*š́*	𐬳
𐬴			ż	43		*ṣ̌*	𐬴
𐬲	24	j	ż	45	*ž*	*ž*	𐬲
𐬵	19	h	h	46	*h*	*h*	𐬵
[19]	—	—	—	—	—		[19]

Table 1: Transcription systems for Avestan

2. Encoding Avestan

2.1 A 7-bit rendering

In the middle of the 1980s, when the project of digitizing the Avestan corpus was initiated,[21] there was no use in trying to encode the texts in the original script, given that the line-based desktop computer available for the project was not programmable to non-Latin scripts.[22] The same holds true for several special characters used in K. HOFFMANN's transliteration system. As a matter of fact, the character inventory usable for the given task consisted of nothing but the items pertaining to the plain ASCII standard,[23] plus a few extra characters necessary for the encoding of German and Skandinavian languages, all stored in the 7-bit range of character encoding (code values of 0 to 127; cf. Table 2 showing the character set of the computer used, with the German non-ASCII characters printed on a shaded background). To maintain the principle of a unique one-to-one rendering of (transliterated) Avestan characters, the existing inventory had to be applied with awkward-seeming but 'natural' equivalences such as $ = *š́*, ö = *ə*, or Z = *ž*. The transliteration system thus arrived at differed greatly from that of comparable digitization projects such as that of the Ṛgveda Saṃhitā,[24] the most ancient text collection of Vedic Sanskrit, which made ample use of digraphical and trigraphical combinations of ASCII characters (cf. the example in Table 3).[25] The advantage of the 'clumsy' one-to-one encoding of (transcriptional) Avestan simply consisted in the fact that it could easily be converted into any other code, without any further consideration of the length of coherent character sequences; in addition, we may state that the inventory necessary for rendering Vedic Sanskrit was much larger than that covered by Avestan (because of the great number of accented vowels it has to cover), and a 7-bit-based one-to-one rendering system (which cannot provide code points for more than ca. 120 characters) would not have been applicable for it.[26] Another reason to stick to a one-byte representation lay in the fact that the amount of disk space available was extremely limited when the Avesta project was started; there was no hard disk available yet, and the ca. 1.2 Million characters of the text collection were just what the two floppy disks manageable by the system could store (in a database application that had to be programmed especially for this task).

	0										1									
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000	◄	¿	X̄	N̄	α	β	Σ	Δ	←	σ	↑	λ	μ		τ	π	θ	Æ	æ	Å
020	å	Ä	ä	Ö	ö	Ü	ü	├	│	─	£	┼		!	“	#	$	%	&	‘
040	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;
060	<	=	>	?	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
080	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_	`	a	b	c
100	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w
120	x	y	z	{	|	}	~	░												
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
	0										1									

Table 2: 7-bit character set applied in 1985

R700123011	AGNI!M+ NA!RO DI:!D)ITIB)IR ARA!N\YOR HA!STACYUTI: JANAYANTA PRAS=ASTA
R700123012	!M / DU:RED9!S=AM+ G9HA!PATIM AT)ARYU!M
R700123021	TA!M AGNI!M A!STE VA!SAVO NY 9&N\VAN SUPRATICA!KS\AM A!VASE KU!TAS= CI
R700123022	T / DAKS\A:!YYO YO! DA!MA A:!SA NI!TYAH-
R700123031	PRE!DD)O AGNE DI:DIHI PURO! NO! 'JASRAYA: SU:RMYA:& YAVIS\T\)A / TVA:!
R700123032	M+ S=A!S=VANTA U!PA YANTI VA:!JA:H-
1	*agníṃ náro dī́dhitibhir aráṇyor hástacyutī janayanta praśastám / dūredṛ́śaṃ gṛhápatim atharyúm*
2	*tám agním áste vásavo ny ṛ̀ṇvan supraticákṣam ávase kútaś cit / dakṣā́yyo yó dáma ā́sa nítyaḥ*
3	*préddho agne dīdihi puró nó 'jasrayā sūrmyā̀ yaviṣṭha / tvā́ṃ śáśvanta úpa yanti vā́jāḥ*

Table 3: Encoding of the Texas R̥gveda (7,1,1-3) contrasted with the usual transcription

2.2 An 8-bit rendering

After having moved to an IBM-DOS-based system in 1986, the transcriptional data could for the first time be visualized both on a printer and on the screen. Equipped with a programmable EGA[27] graphics card and a 70 MB hard disk, the IBM-compatible PC used then was much better suited to the task of completing the electronic corpus of Avestan. Software for entering larger specimens of non-conventional text in a structured way was also available by then: even though it was still line-based, WordPerfect 4.1 was an excellent basis for this task as it enabled the user to check his or her input in a “Reveal Codes” screen and provided an interface for rendering the special transcription characters correctly even on a Laser Printer. For the encoding of Avestan (and other ancient Indo-European languages) in WP 4.1, a special font could thus be designed for both screen and printer representation; different from the 7-bit font used before, this was 8-bit based, with all “special” (non-ASCII) characters stored in the “upper” character range (code values extending from 128 to 255, cf. Table 4).

	0										1									
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000		̣	˙	¯	˘	ˇ	´	`	¨	^	˝	̣	˙	̬	˘	ˇ	´	`	¨	‶
020	″	§	⌒	˛	¸	Ł	Þ	^{h}	^{u}	°	ʻ	ʿ		!	“	#	†	°	$^{+}$	'
040	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;
060	<	=	>	?	√	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
080	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	¯	`	a	b	c
100	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w
120	x	y	z	{	|	}	~	≈	∵	ü	é	â	ä	à	å	ç	ê	ë	è	ï
140	î	ì	Ä	ø	ė	æ	œ	ô	ö	ò	û	ù	ẏ	Ö	Ü	ã	ẽ	ĩ	õ	ũ
160	á	í	ó	ú	ñ	ŋ	ā	ē	ī	ō	ū	ā́	ǰ	ī́	ł	ū́	ā̀	ĕ	ī̀	ı
180	ū̀	ą̇	ā̊	x́	x^{v}	ž	ŋ^{u}	ṛ̀	ĭ	r̄	ŭ	ą	ę	į̇	ǫ	ų	i̯	u̯	ə	ə̄
200	ə̯	ą̃	ą́	ę̃	ę́	ė̃	ė́	į̇̃	į̇́	ų̃	ų́	ū̃	ỹ	ý	β	ƀ	č	ḍ	đ	δ
220	ǵ	ġ	g̰	γ	ḫ	ß	ḥ	ƕ	ḱ	ḷ	ḷ́	ḹ	l̃	ṃ	m̃	m̊	m̨	ṅ	ń	ŋ́
240	ṇ	ṛ	ṛ́	ṝ	ṝ́	r̃	ś	ṣ	š	š́	ṣ̌	t̰	ṭ	ϑ	þ					
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
	0										1									

Table 4: 8-bit font used for the encoding of Avestan and other ancient Indo-European languages (1986-1989)

2.3 The first 16-bit encoding scheme

In 1988, the project proceeded one step further by applying the first 16-bit character encoding available for PCs. With WordPerfect 5.0, the user had at hand a total of 1632 uniquely encodable characters, among them Greek, Cyrillic, Hebrew, and Japanese (*hiragana* and *katakana*) sets, but also a large set of Latin characters with diacritics that were not covered by 7-bit ASCII or its 'western' 8-bit successor, the ANSI standard.[28] For the encoding of the 'idiosyncratic' transcription of Avestan, even this character 'block' (cf. Table 5) was not sufficient though; instead, the project had to rely upon the extra-block of 'user definable' entities, which comprised up to 255 additional items, and characters such as *ẏ, ą̇* or *ə̄* had to be assigned code points in that range (cf. Table 6). For the screen rendering, which was still line-based, WP 5.0 provided a sophisticated solution to extend the 8-bit-based character set of the graphics cards used in PCs to 512 characters, and this was programmable to display the extra characters of Avestan transcription.

No.	Character	No.	Character	No.	Character
1,23	*ß*	1,90	*Ă*	1,162	*Ő*
1,24	*ı*	1,91	*ă*	1,163	*ő*
1,25	*ȷ*	1,92	*Ā*	1,164	*Ō*
1,26	*Á*	1,93	*ā*	1,165	*ō*
1,27	*á*	1,94	*Ą*	1,166	*Œ*
1,28	*Â*	1,95	*ą*	1,167	*œ*
1,29	*â*	1,96	*Ć*	1,168	*Ŕ*
1,30	*Ä*	1,97	*ć*	1,169	*ŕ*
1,31	*ä*	1,98	*Č*	1,170	*Ř*
1,32	*À*	1,99	*č*	1,171	*ř*
1,33	*à*	1,100	*Ĉ*	1,172	*Ŗ*
1,34	*Å*	1,101	*ĉ*	1,173	*ŗ*
1,35	*å*	1,102	*Ċ*	1,174	*Ś*
1,36	*Æ*	1,103	*ċ*	1,175	*ś*
1,37	*æ*	1,104	*Ď*	1,176	*Š*
1,38	*Ç*	1,105	*ď*	1,177	*š*
1,39	*ç*	1,106	*Ĕ*	1,178	*Ş*
1,40	*É*	1,107	*ĕ*	1,179	*ş*
1,41	*é*	1,108	*Ė*	1,180	*Ŝ*
1,42	*Ê*	1,109	*ė*	1,181	*ŝ*
1,43	*ê*	1,110	*Ē*	1,182	*Ť*
1,44	*Ë*	1,111	*ē*	1,183	*ť*
1,45	*ë*	1,112	*Ę*	1,184	*Ţ*
1,46	*È*	1,113	*ę*	1,185	*ţ*
1,47	*è*	1,114	*Ǵ*	1,186	*Ŧ*
1,48	*Í*	1,115	*ǵ*	1,187	*ŧ*
1,49	*í*	1,116	*Ğ*	1,188	*Ŭ*
1,50	*Î*	1,117	*ğ*	1,189	*ŭ*
1,51	*î*	1,118	*Ǧ*	1,190	*Ű*
1,52	*Ï*	1,119	*ǧ*	1,191	*ű*
1,53	*ï*	1,120	*Ģ*	1,192	*Ū*
1,54	*Ì*	1,121	*ģ*	1,193	*ū*
1,55	*ì*	1,122	*Ĝ*	1,194	*Ų*
1,56	*Ñ*	1,123	*ĝ*	1,195	*ų*
1,57	*ñ*	1,124	*Ġ*	1,196	*Ů*
1,58	*Ó*	1,125	*ġ*	1,197	*ů*
1,59	*ó*	1,126	*Ĥ*	1,198	*Ũ*
1,60	*Ô*	1,127	*ĥ*	1,199	*ũ*
1,61	*ô*	1,128	*Ħ*	1,200	*Ŵ*
1,62	*Ö*	1,129	*ħ*	1,201	*ŵ*
1,63	*ö*	1,130	*İ*	1,202	*Ŷ*
1,64	*Ò*	1,131	*i*	1,203	*ŷ*
1,65	*ò*	1,132	*Ī*	1,204	*Ź*

No.	Character	No.	Character	No.	Character
1,66	*Ú*	1,133	*ī*	1,205	*ź*
1,67	*ú*	1,134	*Į*	1,206	*Ž*
1,68	*Û*	1,135	*į*	1,207	*ž*
1,69	*û*	1,136	*Ĩ*	1,208	*Ż*
1,70	*Ü*	1,137	*ĩ*	1,209	*ż*
1,71	*ü*	1,138	*Ĳ*	1,210	*Ŋ*
1,72	*Ù*	1,139	*ĳ*	1,211	*ŋ*
1,73	*ù*	1,140	*Ĵ*	1,212	*D̄*
1,74	*Ÿ*	1,141	*ĵ*	1,213	*d̄*
1,75	*ÿ*	1,142	*Ķ*	1,214	*L̄*
1,76	*Ã*	1,143	*ķ*	1,215	*l̄*
1,77	*ã*	1,144	*Ĺ*	1,216	*N̄*
1,78	*Đ*	1,145	*ŀ*	1,217	*n̄*
1,79	*đ*	1,146	*Ľ*	1,218	*R̀*
1,80	*Ø*	1,147	*ľ*	1,219	*r̀*
1,81	*ø*	1,148	*Ļ*	1,220	*S̄*
1,82	*Õ*	1,149	*ļ*	1,221	*s̄*
1,83	*õ*	1,150	*Ŀ*	1,222	*T̄*
1,84	*Ý*	1,151	*ŀ*	1,223	*t̄*
1,85	*ý*	1,152	*Ł*	1,224	*Y̆*
1,86	*Ð*	1,153	*ł*	1,225	*y̆*
1,87	*ð*	1,154	*Ń*	1,226	*Ỳ*
1,88	*Þ*	1,155	*ń*	1,227	*ỳ*
1,89	*þ*	1,156	*ʼN*	1,228	*D̕*
		1,157	*ŉ*	1,229	*d̕*
		1,158	*Ň*	1,230	*Ơ*
		1,159	*ň*	1,231	*ơ*
		1,160	*Ņ*	1,232	*Ư*
		1,161	*ņ*	1,233	*ư*

Table 5: The ‘Latin Extended’ Block of WP 5.0

No.	Character	No.	Character	No.	Character
12,0	*ẽ*	12,40	*š́*	12,76	*ṣ*
12,1	*ą̃*	12,41	*t̰*	12,77	*ṭ*
12,2	*ę̃*	12,42	*i̯*	12,78	*ṯ*
12,3	*į̄̃*	12,43	*u̯*	12,79	*ẓ*
12,4	*ų̃*	12,44	*B̠*	12,84	*ẏ*
12,5	*ū̃*	12,45	*D̠*	12,85	*ā́*
12,6	*ū́*	12,46	*Ḍ*	12,86	*ī́*
12,7	*ą́*	12,47	*Ĕ*	12,87	*ā̀*
12,8	*ę́*	12,48	*Ḡ*	12,88	*ī̀*
12,9	*į́*	12,49	*Ḥ*	12,89	*ū̀*

No.	Character	No.	Character	No.	Character
12,10	*ų́*	12,50	*H̯*	12,90	*ą̇*
12,11	*ī̃*	12,51	*Ĭ*	12,91	*ā̊*
12,12	*m̃*	12,52	*J̌*	12,92	*x́*
12,13	*r̃*	12,53	*Ḵ*	12,93	x^v
12,14	*r̂*	12,54	*Ŏ*	12,94	$ŋ^u$
12,15	*ȑ*	12,55	*P̱*	12,95	*ṛ̀*
12,16	*â*	12,56	*R̄*	12,96	*ə̄*
12,17	*ê*	12,57	*Ṣ*	12,97	*ə̯*
12,18	*î*	12,58	*Ṭ*	12,98	*ƀ*
12,19	*ô*	12,59	*Ṯ*	12,99	*ǥ*
12,20	*û*	12,60	*Ẓ*	12,100	*ƕ*
12,21	*ȁ*	12,61	∵	12,101	*ḱ*
12,22	*ȅ*	12,62	*ḇ*	12,102	*ḷ́*
12,23	*ȉ*	12,63	*ḏ*	12,103	*ḹ*
12,24	*ȍ*	12,64	*ḍ*	12,104	*m̊́*
12,25	*ȕ*	12,65	*ĕ*	12,105	*m̨*
12,26	*ė́*	12,66	*ə*	12,106	*ṅ*
12,27	*ė̃*	12,67	*ḡ*	12,107	*ŋ́*
12,28	*ỹ*	12,68	*ḥ*	12,108	*ṇ*
12,29	*Ẽ*	12,69	*ḫ*	12,109	*ṛ*
12,30	*Ǫ*	12,70	*ĭ*	12,110	*ṛ́*
12,31	*ǫ*	12,71	*ǰ*	12,111	*ṝ*
12,38	*ḷ*	12,72	*ḵ*	12,112	*ṝ́*
12,39	*ṃ*	12,73	*ŏ*	12,113	*ş̌*
		12,74	*p̄*	12,114	*ṃ̄*
		12,75	*r̄*	12,115	*ṇ̄*

Table 6: Assignment of the 'User definable' Block of WP 5.0

2.4 Rendering the original script

The next version of WordPerfect, 5.1, was even programmable to display and handle the Avestan original script with its right-to-left directionality. The prerequisite for this was the installation of either the Hebrew or the Arabic language package, both of which added the necessary functionality for switching between bidirectional text passages. For Avestan, however, the packages offered no code space off-hand; instead, the Avestan characters had to be mapped onto one of the character sets of either Hebrew (block 9) or Arabic (blocks 13 and 14). As the latter was designed to imply the automatic adaptation of letters to their left and right environment (a feature not relevant to Avestan), the Hebrew block was much better suited for this purpose. The resulting assignment is illustrated in Table 7; for lack of demand, it was never applied to the rendering of the corpus.

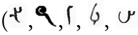
WP	Heb.	Av.	Trs.	WP	Heb.	Av.	Trs.	WP	Heb.	Av.	Trs.
9,0	א	𐬀	*a*	9,19	ף	𐬚	*ϑ*	9,42	·	𐬏	*ū*
9,1	ב	𐬡	*β*	9,20	פ	𐬟	*f*	9,84	בּ	𐬠	*b*
9,2	ג	𐬖	*γ*	9,21	ץ	𐬘	*j*	9,85	גּ	𐬔	*g*
9,3	ד	𐬜	*δ*	9,22	צ	𐬗	*c*	9,87	דּ	𐬛	*d*
9,4	ה	𐬵	*h*	9,23	ק	𐬢	*ŋ*	9,88	וּ	𐬄	*ą*
9,5	ו	𐬬	*v*	9,24	ר	𐬭	*r*	9,89	וֹ	𐬅	*ą̇*
9,6	ז	𐬰	*z*	9,25	שׁ	𐬱	*š*	9,90	זּ	𐬲	*ž*
9,7	ח	𐬑	*x*	9,26	שׂ	𐬴	*ṣ̌*	9,91	הּ	𐬓	*x*^{v}
9,8	ט	𐬝	*ṯ*	9,31	◌ִ	𐬌	*i*	9,92	טּ	𐬜	*δ*$_{2}$
9,9	י	𐬪	*ẏ*	9,32	◌ֵ	𐬉	*ē*	9,93	יּ	𐬫	*y*
9,10	ך	𐬕	*ġ*	9,33	◌ֶ	𐬈	*e*	9,94	כּ	𐬐	*k*
9,11	כ	𐬒	*x́*	9,34	◌ֻ	𐬎	*u*	9,103	מּ	𐬣	*ŋ́*
9,12	ל	𐬮	(*l*)	9,35	◌ָ	𐬁	*ā*	9,104	נּ	𐬧	*ṇ*
9,13	ם	𐬩	*m̨*	9,36	◌ַ	𐬃	*ā̊*	9,106	סּ	𐬯	*s*
9,14	מ	𐬨	*m*	9,37	◌ְ	𐬆	*ə*	9,107	פּ	𐬞	*p*
9,15	ן	𐬦	*ń*	9,38	◌ֲ	𐬋	*ō*	9,108	צּ	𐬢	*ŋ*
9,16	נ	𐬥	*n*	9,39	◌ֱ	𐬍	*ī*	9,109	קּ	𐬤	*ŋ*^{u}
9,17	ס	𐬯	*s*	9,40	◌ֳ	𐬂	*å*	9,111	שּ	𐬳	*ś*
9,18	ע	𐬇	*ə̄*	9,41	·	𐬊	*o*	9,114	תּ	𐬙	*t*

Table 7: Avestan characters mapped onto the Hebrew character set of WP 5.1

2.5 Towards unique encoding: Unicode

With the introduction of the World Wide Web in about 1994, it became necessary to provide a unique encoding scheme for the Avestan texts that was not restricted to proprietary formats. As none of the code pages that were usable in WWW applications then covered the special characters used in the transcription of Avestan, let alone the original Avestan script, the project had to rely upon the emerging 'Unicode' standard right from the beginning even though there was practically no support for this available when the first specimens of the corpus were put online on the server of the TITUS project in 1996. The first sample page, which is still available today (cf. http://titus.uni-frankfurt.de/unicode/samples/homyast.htm), displays in Roman transcription a part of the so-called 'Hōm-Yašt' (i.e. Yasna 9,1-11,8) together with its Middle Persian ('Pahlavī') and Sanskrit translations and liturgical prescriptions in Pāzend (i.e. Middle Persian written in Avestan script). The page clearly indicates what was encodable and retrievable in the early years of Unicode: many characters could not be visualized because they (or their elements, diacritics or basic characters) were not covered by standard fonts, or they had to be left open as there were no code points available for them yet. Meanwhile, 15 years after these first attempts, Unicode has become prevalent as the most widely used encoding standard in the Web, and there is no longer any problem in encoding, retrieving and displaying the transcriptional data of the Hōm-Yašt or any other Avestan text (cf. the online edition in http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/avest010.htm#Avest._Y_9). It is true that many of the combinations of characters with diacritics that are used in the transcription (e.g., *ą̇*, *t̰*, or *x́*) cannot be encoded as such, i.e., as 'precomposed characters', because there are no code points available for them; instead they have to be encoded as sequences of basic characters and diacritics,[29] and it has taken quite some time until 'system fonts' and 'rendering engines' were able to display this kind of combinations in an acceptable way.

As to the original script, the development of a corresponding code block within Unicode took even longer, and only with the publication of Unicode version 5.2.0 in October 2009 has this goal been achieved.[30] The Avestan block consisting of code points 10B00 through 10B3F[31] now enables us for the first time to encode in a standardized way the complete text of the Avesta in the original script. However, for lack of standard fonts that comply to this standard, it will take some more time to make this encoding readily available to the public.

Still, we have to admit that the encoding provided by the new Unicode block is not exhaustive, given that there is still a small set of letters that have not been assigned a code point. The reason is that these letters () have been regarded as mere glyph variants of other characters (𐬅 = ą̇, 𐬜 = δ, 𐬦 = ń, 𐬬 = v, and 𐬵 = h), mostly in accordance with traditional usage which did not provide separate transcriptions for them. However, this decision brings about a dilemma, not only with respect to displaying them in a scholarly context such as the present paper (as a matter of fact, the five letters in question are represented by images here): if we wanted to challenge the assumption of their being functionally identical with their encodable 'partners', we would have to check whether they only occur in distinct environments and never side by side with them within one and the same manuscript. But to check this thoroughly, we would have to encode the texts of all manuscripts accordingly – which we cannot, as there are no code points for these letters available. It is true that provisional code points could be provided in the 'Private Use Area' of Unicode;[32] however, the use of code points of this area may still lead to problems depending on systems and software used, and it is therefore not recommendable. For the time being, I suggest to solve the problem via transliteration, by assigning special (adapted) transliteration symbols to the five letters in question.

3 Summary

The different approaches to the encoding of the Avestan script, first in transliteration and later in the original form, are summarized in Table 8 below. The table also includes the five letters for which no Unicode code points are available, together with a proposal for their transliteration. The combinations of 𐬌𐬌 = ii and 𐬎𐬎 = uu are not included as they have been encoded right from the beginning as sequences of the two characters each that they contain.

Orig.	HP 86B[33]		WP4[33]		WP5[34]		Trl.[35]	Unicode[36]		Orig.
𐬀	a	97	*a*	97	0,97	9,0	*a*	0061	10B00	𐬀
𐬁	A	65	*ā*	166	1,93	9,35	*ā*	0101	10B01	𐬁
𐬂	ü	26	*å*	134	1,35	9,40	*å*	00E5	10B02	𐬂

Orig.	HP 86B[33]		WP4[33]		WP5[34]		Trl.[35]	Unicode[36]		Orig.
𐬃	Ü	25	*ā̊*	182	12,91	9,36	*ā̊*	0101 + 030A	10B03	𐬃
𐬄	ä	22	*ą*	191	1,95	9,88	*ą*	0105	10B04	𐬄
𐬅	Ä	21	*ą̇*	181	12,90	9,89	*ą̇*	0105 + 0307	10B05	𐬅
𐬅							*ą̄*	0105 + 0304		𐬅 [19]
𐬆	ö	24	*ə*	198	12,66	9,37	*ə*	01DD	10B06	𐬆
𐬇	Ö	23	*ə̄*	199	12,96	9,18	*ə̄*	01DD + 0304	10B07	𐬇
𐬈	e	101	*e*	101	0,101	9,33	*e*	0065	10B08	𐬈
𐬉	E	69	*ē*	167	1,111	9,32	*ē*	0113	10B09	𐬉
𐬊	o	111	*o*	111	0,111	9,41	*o*	006F	10B0A	𐬊
𐬋	O	79	*ō*	169	1,165	9,38	*ō*	014D	10B0B	𐬋
𐬌	i	105	*i*	105	0,105	9,31	*i*	0069	10B0C	𐬌
𐬍	I	73	*ī*	168	1,133	9,39	*ī*	012B	10B0D	𐬍
𐬎	u	117	*u*	117	0,117	9,34	*u*	0075	10B0E	𐬎
𐬏	U	85	*ū*	170	1,193	9,42	*ū*	016B	10B0F	𐬏
𐬐	k	107	*k*	107	0,107	9,94	*k*	006B	10B10	𐬐
𐬑	x	120	*x*	120	0,120	9,7	*x*	0078	10B11	𐬑
𐬒	X	88	*x́*	183	12,92	9,11	*x́*	0078 + 0301	10B12	𐬒
𐬓	w	119	x^{v}	184	12,93	9,91	x^{v}	0078 + 036E	10B13	𐬓
𐬔	g	103	*g*	103	0,103	9,85	*g*	0067	10B14	𐬔
𐬕	K	75	*ġ*	221	1,125	9,10	*ġ*	0121	10B15	𐬕
𐬖	G	71	*γ*	223	8,7	9,2	*γ*	03B3	10B16	𐬖
𐬗	c	99	*c*	99	0,99	9,22	*c*	0063	10B17	𐬗
𐬘	j	106	*j*	106	0,106	9,21	*j*	006A	10B18	𐬘
𐬙	t	116	*t*	116	0,116	9,114	*t*	0074	10B19	𐬙
𐬚	F	70	*ϑ*	253	8,17	9,19	*θ*	03B8[37]	10B1A	𐬚
𐬛	d	100	*d*	100	0,100	9,87	*d*	0064	10B1B	𐬛
𐬜	D	68	*δ*	219	8,9	9,3	*δ*	03B4	10B1C	𐬜
𐬜							*δ́*	03B4 + 0301		𐬜
𐬝	T	84	*t̰*	251	12,78	9,8	*t̰*	0074 + 0330	10B1D	𐬝

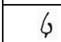
Orig.	HP 86B[33]		WP4[33]		WP5[34]		Trl.[35]	Unicode[36]		Orig.
𐬞	p	112	*p*	112	0,112	9,107	*p*	0070	10B1E	𐬞
𐬟	f	102	*f*	102	0,102	9,20	*f*	0066	10B1F	𐬟
𐬠	b	98	*b*	98	0,98	9,84	*b*	0062	10B20	𐬠
𐬡	B	66	*β*	214	8,3	9,1	*β*	03B2	10B21	𐬡
𐬢	q	113	*ŋ*	165	1,211	9,23	*ŋ*	014B	10B22	𐬢
𐬣	@	64	*ŋ́*	239	12,107	9,103	*ŋ́*	014B + 0301	10B23	𐬣
𐬤	Q	81	*ŋ*^{u}	186	12,94	9,109	*ŋ*^{u}	014B + 0367	10B24	𐬤
𐬥	n	110	*n*	110	0,110	9,16	*n*	006E	10B25	𐬥
𐬦	\	92	*ń*	238	1,155	9,15	*ń*	0144	10B26	𐬦
							ṅ	1E45		
𐬧	N	78	*ṇ*	240	12,108	9,104	*ṇ*	1E47	10B27	𐬧
𐬨	m	109	*m*	109	0,109	9,14	*m*	006D	10B28	𐬨
𐬩	M	77	*m̨*	77	12,105	9,13	*m̨*	006D + 0328	10B29	𐬩
𐬪	Y	89	*ẏ*	152	12,84	9,9	*ẏ*	1E8F	10B2A	𐬪
𐬫	y	121	*y*	121	0,121	9,93	*y*	0079	10B2B	𐬫
𐬬	v	118	*v*	118	0,118	9,5	*v*	0076	10B2C	𐬬
							v̇	0076 + 0307		
𐬭	r	114	*r*	114	0,114	9,24	*r*	0072	10B2D	𐬭
𐬮						9,12	*l*	006C	10B2E	𐬮
𐬯	s	115	*s*	115	0,115	9,17	*s*	0073	10B2F	𐬯
𐬰	z	122	*z*	122	0,122	9,6	*z*	007A	10B30	𐬰
𐬱	S	83	*š*	248	1,117	9,25	*š*	0161	10B31	𐬱
𐬲	C	67	*š́*	249	12,40	9,111	*š́*	0161 + 0301	10B32	𐬲
𐬳	$	36	*ṣ̌*	250	12,113	9,26	*ṣ̌*	0161 + 0323	10B33	𐬳
𐬴	Z	90	*ž*	185	1,207	9,90	*ž*	017E	10B34	𐬴
𐬵	h	104	*h*	104	0,104	9,4	*h*	0068	10B35	𐬵
							h́	0068 + 0301		
·	.	46	.	46	0,46	0,46	.	002E	2E31	⸱
𐬻	:	58	∵	128	12,61	12,61	∵	2235	10B3B	𐬻

Table 8: Encodings used for the Avestan script (original and transliteration)[38]

[1] I.e., from the middle of the sixth century B.C. up to the middle of the eighth century A.D.

[2] Cf. GIPPERT (1995).

[3] http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/avest.htm.

[4] http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/yasna/yasnavar/yasna.htm.

[5] Most of the text was input by SONJA FRITZ in 1985–88, the necessary programming being undertaken by the present author. Additions were provided by H. KUMAMOTO, M. DE VAAN and others.

[6] This is GELDNER's *Avesta* (1889-96). The same is true for WESTERGAARD (1852-54).

[7] Cf. JEAN CHARDIN (1711: 108–9 with Table LXX (Fig. T)); the first edition of CHARDIN's travel report, of which one volume appeared in 1686 in English (second edition 1691) and French, does not contain any relevant information (the subsequent volumes seem not to have been published then). Previous accounts of the Parsees and their religion are the second volume of HENRY LORD (1630), *A Display of two forraigne sects in the East Indies*, published under the title *The Religion of the Persees* etc. (see bibliography), and GABRIEL DE CHINON (1671); they mention the existence of the script without going into details (LORD 1630, p. [2 of 'The proeme']: 'I gained the knowledge of what hereafter I shall deliuer, as it was compiled in a booke writ in the *Persian* Character, containing their Scriptures, and in their owne language, called their ZVNDAVASTAVV.'; CHINON (1671, p. 437): '... voyans qu'ils n'avoient plus de Livres, en écrivirent un de ce qui leur étoit resté en mémoire de ceux qu'ils avoient tant lûs de fois. Celui-là leur est resté, je l'ai vû, il est assez gros, & écrit en caractères fort differens du Persan, de l'Arabe, & des autres Langues du païs, & et qui leurs sont particulieres. Ils le sçavent lire, mais ils disent qu'ils ne l'entendent pas.')

[8] CHARDIN (1711, Vol. 9, 109): 'J'ai inseré dans cet ouvrage, pour la satisfaction des Curieux, un Alphabet de ces anciens *Perses*, ou *Guebres*'.

[9] The illustration is taken from the online version provided by Google Books of the copy of vol. 9 kept in the Lyon Public Library (http://books.google.de/books?id=IjBhi4sF2loC); unfortunately, the Table has been clipped dramatically in the reproduction so that it shows only parts of two lines. The reproduction of the copy of the Bayerische Staatsbibliothek in http://www.mdz-nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-bsb10620739-2 (cf. also http://books.google.de/books?id=2HpCAAAAcAAJ) is even worse in this respect.

[10] An even shorter notice is contained in vol. 5 of CHARDIN's *Voyage* (1711, p. 41).

[11] CHARDIN (1711, ib.): 'en grandes & petites Lettres'.

[12] HYDE (o.c., p. 427): 'Lingua Péhlavi, quae verè Persica'.

[13] Not all entries are comprehensible.

[14] 'Literae in Libris ZEND & PAZEND, juxta Apographum D. *Hyde*, usitatae; una cum Ligaturis & Abbreuiaturis Zendicis', Table added after p. 580.

[15] ANQUETIL-DUPERRON (1771: Pl. VIII, opposite p. 424 ; in other bindings, opposite p. 432).
[16] PIETRASZEWSKI (1858: XI and 1).
[17] A corresponding table is printed in HOFFMANN/FORSSMAN (1996: 41).
[18] The differences are explicitly summarized in HOFFMANN/FORSSMAN (o.c., p. 43).
[19] Character not assigned a Unicode code point (cf. below).
[20] This is not a single character but the combination of two identical ones.
[21] The digitization was undertaken in connection with the project of a new dictionary of the Avestan language (run by B. SCHLERATH in Berlin), which was funded by the DFG in 1985-1988 with the exception of the computer equipment that had to be purchased for it.
[22] The computer used was a Hewlett-Packard 86B; cf., e.g., http://www.hpmuseum.net/display_item.php?hw=35. The restriction concerned the screen but not programmable printers such as the 24-dot matrix printer EPSON LQ-1500 used then.
[23] "American Standard Code for Information Interchange".
[24] Project undertaken in the 1970s under the guidance of W.P. LEHMANN by H.S. ANANTHANARAYANA in Austin, Texas.
[25] It must be stated that such a system does meet the requirements of an unambiguous encoding, but not on the basis of one-to-one correspondences.
[26] A similar rendering system is the so-called 'Beta-Code', which was invented for the rendering of Ancient Greek and which has been used until the present day in the 'Thesaurus Linguae Graecae' project at the University of California (Irvine) and in the 'Perseus' project of Tufts University.
[27] "Enhanced Graphics Adapter", a colour graphics card using a programmable 8-bit character set with a dot-matrix of 14 × 8 dots.
[28] "American National Standards Institute"; the standard is also known as ISO standard no. 8859-1.
[29] Cf. chaps. 5.6 ('Normalization') and 2.11 ('Combining characters') of the current Unicode Standard reference, http://www.unicode.org/versions/Unicode6.0.0/ch05.pdf and /ch02.pdf. The reason is that the Unicode Consortium stopped the integration of precomposed characters early in the process of development of the standard. Cf. http://unicode.org/faq/char_combmark.html as to useless attempts to have new precomposed characters added to Unicode.
[30] The proposal, which was prepared by M. Everson and R. Pournader, dates from March 22, 2007 (cf. http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3197.pdf); cf. http://www.unicode.org/Public/UNIDATA/DerivedAge.txt for a full account of when a given code point was first assigned in Unicode.
[31] Cf. http://www.unicode.org/charts/PDF/U10B00.pdf. The block is contained in 'Plane 1', the so-called 'Supplementary Multilingual Plane (SMP)' of Unicode, which is

mostly designed to comprise 'historical' scripts that are no longer used (cf. chap. 2.9 of the current Unicode Standard reference, http://www.unicode.org/versions/Unicode6.0.0/ch02.pdf); in the case of Avestan, this remains problematic as the Parsee communities in India or elsewhere may return to employing the Avestan script one day.

[32] The 'PUA' ranges from E000 to F8FF and comprises up to 6,400 privately definable and usable characters.

[33] Code numbers represent byte values.

[34] Code numbers indicate the block and the number of the character within the WordPerfect character set (left column: transcription; right column: original script mapped onto Hebrew). Block 8 is the Greek character block.

[35] Extended transliteration including the non-encodable letters.

[36] Code points in hexadecimal notation (left column: transcription; right column: original script). Precomposed characters are indicated wherever they exist.

[37] For Greek *thēta*, there are two code points available (03B8 and 03D1); the first one is chosen here as this is associated with the round-shaped variant displayed here while the second is reserved for the 'open' variant (ϑ).

[38] Proposed encodings (for additional transcriptional characters) are marked with a shaded background.