Achtung!
Dies ist eine Internet-Sonderausgabe des Kapitels
„Digital approaches to oriental manuscript studies“
von Jost Gippert (2014).
Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
Comparative Oriental Manuscript Studies. An Introduction
Hamburg: COMSt 2015, 12–30
zu entnehmen.

Attention!
This is a special internet edition of the chapter
“Digital approaches to oriental manuscript studies”
by Jost Gippert (2014).
It should not be quoted as such. For quotations, please refer to the original edition in
Comparative Oriental Manuscript Studies. An Introduction
Hamburg: COMSt 2015, 12–30.

2. Digital and scientific approaches to oriental manuscript studies (JG–IR–FA)

2.1. Digital approaches to oriental manuscript studies (JG)

With the spread of personal computers in the 1980s and early 1990s, studies concerning manuscripts and their contents started to change in both their aims and their methods, and the 'digital turn' has meanwhile embraced nearly all relevant fields. It seems therefore appropriate first to outline the essentials of digital approaches to oriental manuscript studies here; more detailed treatments will be found in the individual chapters following. The present survey focuses on questions of the representation of different scripts (original and transcriptional) and the encoding of characters; the conception of electronic texts, their structuring and their processing; the arrangement of databases, their layout and their handling; and the basics of digital imaging including special relevant methods of photography.

2.1.1. The representation of oriental scripts and the encoding of characters

In the early times of the digital age, attempts to store and process data in oriental languages were for many years hampered dramatically by the fact that computers were not yet able to deal with scripts other than Latin, and even the correct treatment of extra characters such as the 'umlaut vowels' of German or the accented letters of French was anything but guaranteed. The reason was that in a digital environment, the encoding of written text must be based on a given set of correspondences of characters with numerical values, every character being represented by one unique value. To encode the two times 26 letters (lower and upper case) of the Latin alphabet plus the digits from 0 to 9, the punctuation marks, parentheses, and the like, a set of less than 100 unique values is necessary, and this is why the 'stone age' mainframe computers of the 1960s to 1970s were based on a so-called 7-bit encoding: with 7 bits, $2^7 = 128$ characters can be encoded uniquely. The most popular standard developed on this basis is the so-called ASCII standard ('American Standard Code for Information Interchange', see Table 0.2.1), which prevailed in the first personal computers.

It is clear that on the basis of this encoding scheme, English texts could easily be digitized, but German, French, or Spanish texts could not, let alone Greek, Russian, or Arabic texts in their original scripts. This does not mean, however, that it was impossible then to process texts in more 'exotic' languages. What was necessary was the invention of encoding schemes that used more than one 'code point' to represent certain characters. One such scheme, the so-called 'BETA-Code', was applied to encode the ancient Greek texts that are comprised in the 'Thesaurus Linguae Graecae' (TLG), a huge database attempting to cover the complete textual heritage from Homer down to the Middle Ages. Cf. Table 0.2.2 which shows the 7-bit adaptation of the beginning of Hesiod's *Theogony*, contrasted with the 'traditional' rendering in Greek script. It is clear that the 7-bit encoding had at least two disadvantages: it was hardly possible to visualize the text as it should be on a computer screen, and the encoding was not transparent (or 'self-explaining') in the sense that the individual items (letters, diacritics, accent marks) could be easily determined by people who were not involved in the encoding process themselves. It is true that this encoding met the condition of being consistent in that a given sequence of codes always represented the same character, and this is why these texts can be used and analysed even today (and the TLG website still supports it); however, it will be clear that it remains clumsy and hard to handle.

With the extension of the ASCII encoding basis to 8 bits, this problem was at least partially overcome. On an 8-bit (= 1-byte) basis, $2^8 = 256$ characters can be encoded uniquely, and since the early 1980s, many 8-bit encoding schemes were developed and applied, adding 'special' characters such as those representing German *ä*, *ö*, *ü*, the accented vowels *é*, *à*, *ô*, etc. of French, or the Spanish palatal nasal *ñ* to the inventory. Unfortunately, this was not done in an equal, 'standardized' way right from the beginning; instead, several leading computer companies developed their own individual schemes, which resulted in serious problems whenever data were to be exchanged between systems. Tables 0.2.3−5 show the encoding systems used in IBM/DOS computers, Mac computers, and MS-Windows—only the latter one is more or less identical with the 8-bit standard used in many applications up till now, the ANSI standard ('American National Standards Institute') also known as ISO standard no. 8859-1 (the special MS-Windows characters are displayed on a grey background within Table 0.2.5).

Still, these encoding systems were not sufficient for the immediate encoding of other scripts such as Greek, Cyrillic, or Chinese. This is why from the middle of the 1980s on, so-called 'code pages' were

Table 0.2.1 ASCII encoding standard (7-bit)

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000																				
020														!	"	#	$	%	&	'
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

Table 0.2.2 Greek text with its BETA-Code representation (Hesiod, *Theogony*)

```
*MOUSA/WN *(ELIKWNIA/DWN A)RXW/MEQ' A)EI/DEIN,
AI(/ Q' *(ELIKW=NOS E)/XOUSIN O)/ROS ME/GA TE ZA/QEO/N TE,
KAI/ TE PERI\ KRH/NHN I)OEIDE/A PO/SS' A(PALOI=SIN
O)RXEU=NTAI KAI\ BWMO\N E)RISQENE/OS *KRONI/WNOS:
KAI/ TE LOESSA/MENAI TE/RENA XRO/A *PERMHSSOI=O
H)' *(/IPPOU KRH/NHS H)' *)OLMEIOU= ZAQE/OIO
A)KROTA/TW| *(ELIKW=NI XOROU\S E)NEPOIH/SANTO,
KALOU\S I(MERO/ENTAS, E)PERRW/SANTO DE\ POSSI/N.
E)/NQEN A)PORNU/MENAI KEKALUMME/NAI H)E/RI POLLW=|
E)NNU/XIAI STEI=XON PERIKALLE/A O)/SSAN I(EI=SAI,
```

Μουσάων ῾Ελικωνιάδων ἀρχώμεθ' ἀείδειν,
αἵ θ' ῾Ελικῶνος ἔχουσιν ὄρος μέγα τε ζάθεόν τε,
καί τε περὶ κρήνην ἰοειδέα πόσσ' ἁπαλοῖσιν
ὀρχεῦνται καὶ βωμὸν ἐρισθενέος Κρονίωνος·
καί τε λοεσσάμεναι τέρενα χρόα Περμησσοῖο
ἠ' ῎Ιππου κρήνης ἠ' ᾽Ολμειοῦ ζαθέοιο
ἀκροτάτῳ ῾Ελικῶνι χοροὺς ἐνεποιήσαντο,
καλοὺς ἱμερόεντας, ἐπερρώσαντο δὲ ποσσίν.
ἔνθεν ἀπορνύμεναι κεκαλυμμέναι ἠέρι πολλῷ
ἐννύχιαι στεῖχον περικαλλέα ὄσσαν ἰεῖσαι, …

Table 0.2.3 Non-standard 8-bit encoding (‘DOS/IBM’, ‘Extended ASCII’, ‘Codepage 437’)

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000		☺	☻	♥	♦	♣	♠	•	◘	○	◙	♂	♀	♪	♫	☼	►	◄	↕	‼
020	¶	§	▬	↨	↑	↓	→	←	∟	↔	▲	▼		!	"	#	$	%	&	'
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	{	|	}	~	⌂	Ç	ü	é	â	ä	à	å	ç	ê	ë	è	ï
140	î	ì	Ä	Å	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	¢	£	¥	₱	ƒ
160	á	í	ó	ú	ñ	Ñ	ª	º	¿	⌐	¬	½	¼	¡	«	»	░	▒	▓	|
180	┤	╡	╢	╖	╕	╣	║	╗	╝	╜	╛	┐	└	┴	┬	├	─	┼	╞	╟
200	╚	╔	╩	╦	╠	═	╬	╧	╨	╤	╥	╙	╘	╒	╓	╫	╪	┘	┌	█
220	▄	▌	▐	▀	α	ß	Γ	π	Σ	σ	µ	τ	Φ	Θ	Ω	δ	∞	∅	∈	∩
240	≡	±	≥	≤	⌠	⌡	÷	≈	°	∙	·	√	ⁿ	²	■					
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9

Table 0.2.4 Non-standard 8-bit encoding (Mac OS)

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000																				
020														!	“	#	$	%	&	´
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	{	|	}	~		Ä	Å	Ç	É	Ñ	Ö	Ü	á	à	â	ä	ã
140	å	ç	é	è	ê	ë	í	ì	î	ï	ñ	ó	ò	ô	ö	õ	ú	ù	û	ü
160	†	°	¢	£	§	•	¶	ß	®	©	™	´	¨	≠	Æ	Ø	∞	±	≤	≥
180	¥	µ	∂	Σ	Π	π	∫	ª	º	Ω	æ	ø	¿	¡	¬	√	ƒ	≈	Δ	«
200	»	…		À	Ã	Õ	Œ	œ	–	—	"	“	'	’	÷	◊	ÿ	Ÿ	⁄	¤
220	‹	›	fi	fl	‡	•	‚	„	‰	Â	Ê	Á	Ë	È	Í	Î	Ï	Ì	Ó	Ô
240		Ò	Ú	Û	Ù	ı	ˆ	˜	¯	˘	˙	˚	¸	˝	˛	ˇ				
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9

Table 0.2.5 Standardized 8-bit encoding (ANSI / ISO 8859-1 plus MS-Windows / Codepage 1252)

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000																				
020														!	“	#	$	%	&	´
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	{	|	}	~				,	ƒ	„	…	†	‡	ˆ	‰	Š	‹
140	Œ					‘	’	“	“	○	—	–	˜	™	š	›	œ			Ÿ
160		¡	¢	£	☼	¥	|	§	¨	©	ª	«	¬	-	®	¯	°	±	²	²
180	’	µ	¶	-	¸	¹	º	»	¼	½	¾	¿	À	Á	Â	Ã	Ä	Å	Æ	Ç
200	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û
220	Ü	Ý	Þ	ß	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
240	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ				
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9

developed for 8-bit based computers, in which, just as in the examples shown above, the ‘upper’ area exceeding the basic ASCII plain (values above 128) was used to encode various other character sets. Some of these code pages have been standardized within the ISO standard 8859 (see, for example, Table 0.2.6 contrasting the Cyrillic code page ISO 8859-5 with the ANSI standard, ISO 8859-1), and some of them are still used in web pages.

Apart from these ‘official’ extensions, an unknown amount of local or even personal 8-bit encoding systems were developed in the 1980s and 1990s to meet the needs of philologists dealing with oriental languages. As a matter of fact, whenever someone developed and applied a certain font, the encoding of which did not match one of the standardized code pages, a new encoding system was created from scratch. Applying the method of ‘font mapping’, one could thus meet, for example, the requirements of Ancient (‘Polytonic’) Greek to be noted in original characters as well as Iranian languages to be rendered in a scholarly Latin transcription (see Tables 0.2.7−8).

Table 0.2.6 Standardized 8-bit mapping: ISO 8859-1 vs. ISO 8859-5

ISO 8859-1

32		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	47
48	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	63
64	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	79
80	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_	95
96	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	111
112	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~		127
160		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯	175
176	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿	191
192	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	207
208	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß	223
224	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	239
240	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ	255

ISO 8859-5

32		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	47
48	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	63
64	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	79
80	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_	95
96	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	111
112	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~		127
160		Ё	Ђ	Ѓ	Є	S	I	Ї	J	Љ	Њ	Ћ	Ќ	·	Ў	Џ	175
176	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	191
192	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	207
208	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	223
224	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я	239
240	№	ё	ђ	ѓ	є	s	i	ї	j	љ	њ	ћ	ќ	§	ў	џ	255

Table 0.2.7 Non-standard 8-bit encoding: Ancient ('polytonic') Greek

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000		̀	̆	̈	́	̔	̃	·		∘										
020		§			̓	̓	́	̀	́	F	Ϛ	ϗ		!	"	̔	́	̀	̃	'
040	(	)	*	†	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;
060	ή	ῆ	ὴ	?	ς	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	[	ὴ	]	ῆ	̇	̀	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	ή	|	ὴ	ῆ	ᾶ	ᾰ	ü	έ	ὰ	ä	ὰ	ᾶ	ῖ	ε̃	ĕ	ὲ	ï
140	ῖ	ὶ	Ä	ῐ	ŏ	ῠ	ε̃	õ	ö	ὸ	ῦ	ὺ	ϋ	Ö	Ü	ἀ	ἐ	ἰ	ὀ	ὐ
160	ά	ί	ó	ύ	ᾠ	ᾐ	ᾧ	ᾐ	ᾠ	ᾧ	ᾗ	ῐ	ῠ	ἄ	ᾔ	ᾗ	Γ	Δ	ᾔ	ᾗ
180	ᾗ	ᾔ	Θ	ώ	ὼ	Λ	ᾶ	ὼ	Ξ	ώ	Π	ᾰ	Σ	ὼ	ᾤ	Φ	ᾰ	Ψ	Ω	ᾶ
200	ῖ	ῠ	ά	ᾰ	ή	ῆ	ἤ	ὰ	ἐ	ὶ	ὸ	ὺ	ᾳ	ᾰ	ἄ	α	ᾰ	γ	δ	ε
220	ζ	η	ϑ	ι	κ	β	λ	μ	ν	ξ	ᾤ	π	ρ	σ	τ	υ	φ	χ	ψ	ω
240	ϱ	ῖ	ῠ	ᾰ	ᾰ	ῃ	ῳ	ᾳ	ὰ	ἐ	ἰ	ὸ	ὺ	ῦ	ᾰ					
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9

Table 0.2.8 Non-standard 8-bit encoding: Latin font with diacritics for Iranian languages

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
000		̇	̇	̄	̆	̌	́	̀	̈	̂	̋	̇	̇	̌	̆	̌	́	̀	̈	̋
020	″	§	̂	¸	˛	Ł	Þ	ʰ	ᵘ	°	‘	’		¡	“	#	†	°	+	'
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	{	|	}	~	≈	ź	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	ǻ	ǰ	Í	ł	ű	ǟ	ě	Ì	ı
180	ǜ	ą	ǟ	ǵ	xᵘ	ž	ŋᵘ	į	ĭ	ī	ŭ	ą	ę	į	ǫ	ų	į	ų	ə	ā
200	ę	ą̃	ą	ę̃	ę	ě	é	ǐ	́į	ų̃	ų́	ũ	ỹ	ý	β	ƀ	č	ḍ	đ	δ
220	ǵ	ġ	g	γ	ḫ	ß	ḥ	ḫu	ḱ	ļ	Ḯ	Ī	Ì	ṃ	m̃	m̆	ṃ	ṅ	ń	ŋ́
240	ṇ	ṛ	ṛ́	r̄	r̃	ř	ś	ş	š	š	ş	ţ	ṭ	ϑ	þ					
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9

The problem about all this is that whenever 'font mapping' is applied, the basic requirements of consistent encoding, namely the recoverability and exchangeability of data, cannot be guaranteed as there is no unique one-to-one-relation between a character to be encoded and a given digitized value. If, for example, we applied the Greek 8-bit font illustrated in Table 0.2.7, the value of 231 would represent a Greek lower case letter *pi* (π); the same value would stand for a Cyrillic *ča* (ч), however, if we used a font matching the standard codepage ISO 8859-5, and it would represent a Latin *c* with cedilla (*ç*) if we used the plain ANSI standard. This means that whenever an 8-bit encoding was applied in the encoding of tex-

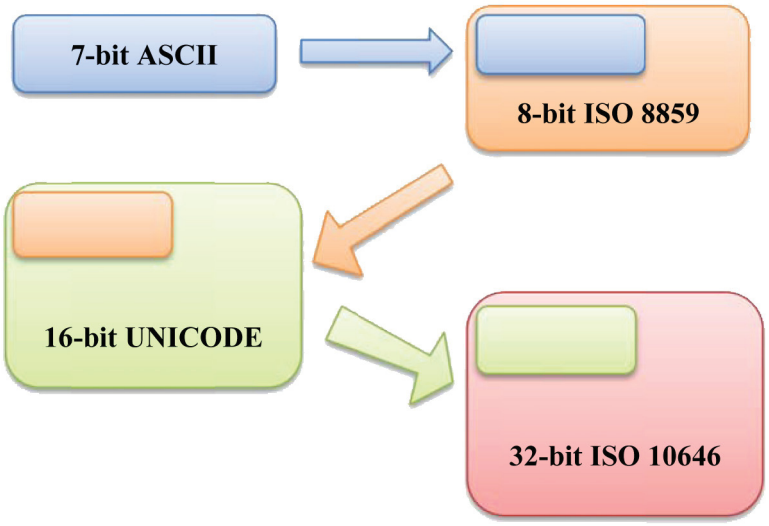


Fig. 0.2.1 From 7-bit to 32-bit encoding

tual materials, additional information had to be stored as to what code page or font encoding was valid for a given character. This information, however, was not encodable as such in a standardized way, being dependent on the idiosyncrasies of word processing programs such as Microsoft Word, and it was lost all too easily when data were transferred across systems. This is all the more true so for scripts with right-to-left direction such as Arabic, which required special encoding solutions in all cases. This is why many textual materials in oriental languages stored electronically in the twentieth century (sometimes even later) in transcribing manuscripts or editing their contents are no longer usable today—or at least hard to process.

To be able uniquely to encode all characters that have been used in writing down human languages including both 'original' scripts and alphabets and linguistic 'transcriptions', the basis of encoding had to be extended far beyond the 1-byte (8-bit) standard. This is exactly what has been undertaken since the early 1990s when the so-called 'Unicode' standard was created: based on 16 bits (or 2 bytes), this standard comprises 2^{16} = 65536 basic 'code points' used for the 'unique' encoding of characters. Considering that for the Chinese script alone, far more than 65,000 different characters have been used throughout history, it is clear that even this standard is not yet sufficient to cover all characters used by mankind at all times. This is why a further extension has been conceived, in the 32-bit standard ISO 10646 which provides a total of (2^{32} =) 4,294,967,296 code points; as a matter of fact, the Unicode standard is but one subset of this near to 'infinite' inventory, just as the ANSI standard (ISO 8859-1) is a subset of Unicode, and the ASCII standard a subset of ANSI (see fig. 0.2.1).

Along with the expansion of the World Wide Web, Unicode encoding has become more and more prominent since the late 1990s, and it is the encoding basis of practically all up-to-date operating systems and word processors today. There can be no doubt that this is a huge advantage for the purposes of oriental manuscript studies. Cf., for example, Table 0.2.9 which shows a few of the 'blocks' of Unicode characters: the distinction of a Cyrillic *ča* (ч) and a Latin *c* with cedilla (*ç*) is now guaranteed by their different code points (hexadecimal number 0447 = decimal 1095 vs. hexadecimal 00E7 = decimal 231), and various Latin-based characters used in transcription systems can now as well be encoded as characters of the Greek, Coptic, or Georgian scripts. In addition, the Unicode standard even comprises information on the directionality of a given character so that Hebrew, Arabic, or Syriac texts can be encoded (and exchanged!) without further programming—provided the system used has implemented the relevant 'blocks' and the rules pertaining to them.

However, even Unicode encoding is not without problems. First of all, it builds upon the so-called character/glyph distinction. According to the definition provided by the Unicode Consortium, a 'glyph is a particular image which represents a character or part of a character', and it 'may have very different shapes' as illustrated by the set of six 'sample glyphs' for the Latin 'character' *a* in Table 0.2.10 (modelled after the diagram in General introduction § 2.1 at <http://www.unicode.org/reports/tr17/tr17-3.html>, accessed March 2014). It will be clear from the example that a 'character', which is what is to be encoded, is an abstraction of all the possible actual forms of a 'letter' that may appear in handwritten or printed

Table 0.2.9 16-bit encoding: Unicode blocks Latin and Cyrillic

Latin

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
000																
001																
002	?	!	“	#	$	%	&	'	(	)	*	+	,	-	.	/
003	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
004	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
005	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
006	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
007	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
008																
009																
00A		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯
00B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
00C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
00D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
00E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
00F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F

Cyrillic

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
040	Ѐ	Ё	Ђ	Ѓ	Є	S	I	Ї	J	Љ	Њ	Ћ	Ќ	Й	Ў	Џ
041	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
042	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
043	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
044	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
045	ѐ	ё	ђ	ѓ	є	s	i	ї	j	љ	њ	ћ	ќ	й	ў	џ
046	Ѡ	ѡ	Ѣ	ѣ	Ю	ю	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ
047	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ	Ѷ	ѷ	Ѹ	ѹ	Ѻ	ѻ	Ѽ	ѽ	Ѿ	ѿ
048	Ҁ	ҁ	҂	҃	҄	҅	҆	҇	҈	҉	Ҋ	ҋ	Ҍ	ҍ	Ҏ	ҏ
049	Ґ	ґ	Ғ	ғ	Ҕ	ҕ	Җ	ж	Ҙ	ҙ	Қ	қ	Ҝ	к	Ҟ	ҟ
04A	Ҡ	ҡ	Ң	ң	Ҥ	ҥ	Ҧ	ҧ	Ҩ	ҩ	Ҫ	ҫ	Ҭ	ҭ	Ү	ү
04B	Ұ	ұ	Ҳ	ҳ	Ҵ	ҵ	Ҷ	ҷ	Ҹ	ҹ	Һ	h	Ҽ	ҽ	Ҿ	ҿ
04C	І	Ӂ	ӂ	Ӄ	ӄ	Ӆ	ӆ	Ӈ	ӈ	Ӊ	ӊ	Ӌ	ӌ	Ӎ	ӎ	Ӏ
04D	Ӑ	ӑ	Ӓ	ӓ	Æ	æ	Ӗ	ӗ	Ә	ә	Ӛ	ӛ	Ӝ	ӝ	Ӟ	ӟ
04E	Ӡ	ӡ	Ӣ	ӣ	Ӥ	ӥ	Ӧ	ӧ	Ө	ө	Ӫ	ӫ	Ӭ	ӭ	Ӯ	ӯ
04F	Ӱ	ӱ	Ӳ	ӳ	Ӵ	ӵ	Ӷ	ӷ	Ӹ	ӹ	Ӻ	ӻ	Ӽ	ӽ	Ӿ	ӿ
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F

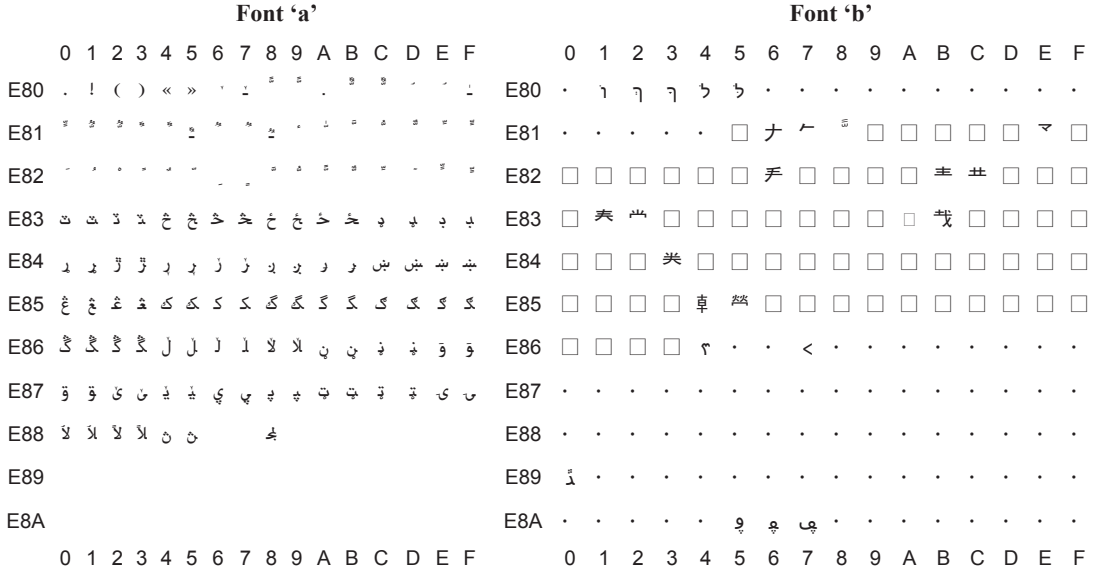
Table 0.2.10 Example of the character/glyph distinction in Unicode

Character	**Sample glyphs**					
a	*a*	*a*	**a**	ɑ	ɐ	a

form, while every single appearance of the letter is regarded as a ‘glyph variant’. This distinction, then, is crucial indeed for manuscript studies, as the assignment of individual ‘letter shapes’ occurring in handwritten sources to ‘abstract’ character values may always be a matter of dispute, especially in a diachronic perspective: we may think, for example, of the emergence of minuscules from majuscules over time, or of ‘new letters’ from former ligatures. As a matter of fact, the decision of the Unicode Consortium to treat the ‘minuscule’ *a* as a character in its own right, with a unique code point, and not to treat all the ‘minuscule’ variants of *a* as glyphs of the one (‘majuscule’) character *A*, which has another code point, may be justified for practical (and traditional) reasons, but it may be problematical indeed for manuscript studies concerning the first millennium. It may be even more problematical when it comes to scripts that are less ‘fixed’ than Latin.

To be sure, the problem of assigning letter forms as appearing in a handwritten context to ‘abstract’ units is not intrinsically determined by digitization, and it is by no means confined to it: just like a scholar of today, who has to decide by what code point he would represent the glyph he ‘reads’ in a manuscript, a scholar using pen or pencil in transcribing a manuscript would have had to decide for an ‘abstract’ character, too, at least when handing his transcript over to a typesetter. There is indeed an important difference, however, in that the purpose of typesetting was limited to a reproduction in print, whereas a digital encoding can be used for other purposes such as automatic indexation as well; here, the consistency of the encoding becomes crucial indeed (cf. below). Another difference concerns the way restrictions could be overcome when necessary, those of a typesetter’s letter case of old and those of an encoding standard of today: the typesetter may have resorted to the production of new types if this was deemed unavoidable (cf. the approaches summarized in the case study on the edition of the Berlin Turfan manuscripts, Ch. 3 § 3.9), and the ‘digital’ scholar, to the tedious process of convincing the Unicode Consortium that a character (not a glyph!) is missing in their standard (cf. the problem of a ‘different letter for *ą* and initial *y*’

Table 0.2.11 16-bit font mapping: The ‘Private Use Area’ of Unicode



in Indian and Iranian manuscripts of the *Avesta*, thematized in case study Ch. 3 § 3.5). Be that as it may, the problem of distinguishing abstract ‘characters’ from ‘glyphs’ as their ‘representations’ is actually one of the history of scripts, their analysis and their usage in general, not of digitization. The development of the Unicode standard has contributed a lot to this question by enforcing thorough investigation, and many of us have been involved in the process of its extension. However, it is a pity that this has often not been determined by scientific reasoning alone but by practical (or even economic) considerations, thus leaving inconsistencies and shortcomings that we still have to cope with.

One such inconsistency lies in the fact that the encoding facilities Unicode provides are not always ‘unique’. This is especially true for the huge amount of combinations of (Latin, Greek, Cyrillic etc.) characters with diacritics it intends to cover, many of which can be encoded ‘as such’, that is as so-called ‘precomposed characters’, or as combinations of the respective ‘basic character’ and the diacritic(s) it carries. For example, the German *ä* can be encoded as the Unicode character no. 226 (U+00E4) or as a sequence of *a* = no. 97 (U+0061) and the ‘umlaut’ diacritic (‘diaeresis’, U+0308); in a similar way, *r* with a macron above and a dot below (*ṝ*) can be encoded as such as no. 7773 (U+1E5D) or as a sequence of *r* (U+0072), macron above (U+0304), and dot below (U+0323), or even as a sequence of *r* with a dot below (*ṛ*, U+1E5B) and a macron above (U+0304). It is true that the different ways of encoding the same ‘composed character’ are essentially equivalent according to the definition of the standard—with the ‘precomposed’ units being considered as the first choice—and should be treated as such by Unicode-based systems; however, users cannot rely upon this in all cases yet, depending on system or software peculiarities.

A similar problem is posed, for example, by Arabic characters, given that Unicode provides code points for both the different ‘surface’ forms they may appear in within words (isolated, final, initial, medial, for example ع, ﻊ, ﻋ, ﻌ; U+FE81 to FEF4) and an ‘idealized’ representation of the underlying ‘abstract’ character (identical in shape with the ‘isolated’ variant) which is meant to be adapted automatically to the context (for example ع, U+062A). Here, too, the different ways of encoding the same character are essentially equivalent according to the definition of the standard, with the ‘idealized’ representations to be used preferably wherever possible.

Another problem that may be crucial in the application of Unicode is the persistence of at least one area that is designed for font mapping. This is the so-called ‘Private Use Area’ (PUA), which comprises 6144 code points for non-predefined characters (in the blocks U+E000–EFFF and F000–F7FF). This area can be assigned *ad libitum* by companies, user groups, or individuals, with the result that additional information is again necessary to distinguish the characters ‘encoded’ in it. Table 0.2.11 shows what can happen when different fonts are applied to visualise PUA encoded characters; in the worst case, the intended information will again be lost. The use of the ‘Private Use Area’ should therefore be avoided wherever possible.

2.1.2. Electronic texts and their structuring

Depending on their envisaged use, electronic texts to be produced and used in oriental manuscript studies require special attention as to their structuring beyond character encoding, too. To clarify what this means, it is helpful to look again at the Greek text we have dealt with above (see Table 0.2.2). Even without any knowledge of the language and script, we will immediately have the impression that this text consists of verses. This is clearly indicated by two signals we are used to in reading poetical texts, namely the relative shortness of lines (with no full justification), and the numbers 1, 5, and 10 attached to the respective lines (in the Greek rendering). There are many further elements of textual structure involved, however. First, we will easily guess that the text consists of several sentences, partially extending across verses and partially consisting of subordinate clauses and phrases: this is indicated by the punctuation marks used. Then, we will be able to state that the text consists of 51 words, in their turn indicated either by empty spaces between them or by punctuation marks adjoining their first or last characters. This may all sound trivial, but as a matter of fact, it can be crucial indeed for textual materials to consider the function of their internal elements and to 'mark them up' accordingly when preparing them for further usage; and this should be done as consistently as the encoding of the characters appearing in words.

So what elements are we talking about? Among the basic elements of every kind of text, we have already mentioned words (consisting of characters when written down), phrases, clauses, and sentences; on a higher level, we will have to deal with sections, paragraphs, chapters, text parts, and the like. For many of these elements, we intuitively adapt signals we have been used to since we were at school, such as spaces indicating word boundaries, full stops indicating sentence breaks, or 'hard' line breaks indicating the end of a section or paragraph. For a consistent encoding of a digital text to be used in a (critical or diplomatic) edition, in an electronic corpus, or for other purposes, this may not be sufficient, though, especially when the contents of oriental manuscripts are concerned. An appropriate example may suffice to illustrate why.

Fig. 0.2.2 shows the upper half of the front fly-leaf of the codex Vienna, ÖNB, Cod.Vind.georg. 2, a Georgian palimpsest manuscript stemming from the Monastery of the Holy Cross at Jerusalem. The leaf in question originally pertained to another codex from the same site, which is kept today in the Dumbarton Oaks Research Library in Washington, DC (MS WAS.1.2), and which represents a menaion covering the months of December to February, starting, in accordance with the Greek Synaxarion, with the commemoration of St Ananias of Persia and SS Onesimus and Solomonus (Solochonus) of Ephesus (see Gippert et al. 2007a, xii–xvii). Even without any knowledge of Georgian, and even neglecting the bad state of preservation especially of the upper part of the page, people experienced with mediaeval manuscripts will easily recognise that there are two different scripts used side by side in it, a majuscule and a minuscule, the former mostly appearing in the four red lines under the ornamental braid of the top, and the latter, mostly in the black text below. A closer look will reveal that even within the black text, there are some red elements, mostly dots accompanying other dots in black, but also some (majuscule) letters (in the fourth line); on the other hand, the first line contains a black letter in a red environment. One might further guess that lines five and ten contain a majuscule letter extending into the left margin, the first in red and the second in black; beyond that, the first text line shows a hanging initial in black, in its turn enclosed by an ornamental structure that might represent another majuscule letter. The colour of the latter is neither red nor black but the same (purple) colour as that of the ornamental heading on the top, and this very colour also appears in an attention mark in the shape of a shaft cross in the left margin; different from the text characters, it is only the contour-lines of these elements that are coloured, not their solid bodies.

As a matter of fact, none of these features is accidental, all of them being related to the meaning and the functions of the textual elements they pertain to. To start with, the four lines in red represent what we might call a heading (actually, it is exactly this use of red ink that has led to the emergence of the word 'rubric'). It begins with the indication of 1 December as the date the following text relates to; the (dative-locative) case form of the month name, *deḳembersa*, appears written in red, while the single character following it in black with an overbar attached to it is the letter *a* in its numerical value, '1', denoting the day of the month. The same letter appearing enclosed at the beginning of the line represents the word-final vowel of *ttuesa*, the word for 'month' in the dative case form corresponding to *deḳembersa*, 'in the month of December'; and its ornamentally-shaped enclosure in violet colour is the word-initial letter of the same word, *t*. The overbar above the *a* here marks the suspension of the characters between *t* and *a* in *t*(*owes*)

Fig. 0.2.2 Vienna, ÖNB, Cod.Vind.georg. 2, front fly-leaf (excerpt)

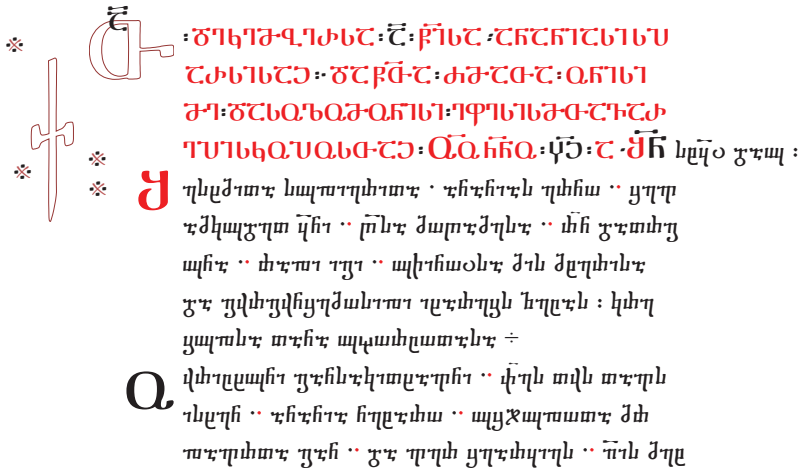
a, not the numerical use of *a* = ‘1’ as in the indication of the day; as a matter of fact, the two overbars seem not to be identical, both being curved a bit differently. Note that between the abbreviated word form and the month name, and also on both sides of the numerical *a* and elsewhere within the rubric, we can detect double dots in black, always used as separators but not necessarily in the same way as a colon (or any other punctuation mark) would be used in modern European languages (including modern Georgian); they simply serve to denote boundaries between major meaningful elements (words, phrases, or clauses).

The text of the rubric continues with the names of the saints commemorated, all in the genitive case as if depending on a head noun like ‘commemoration’: *ç̇midisa : ananiasi s̈parsisay : da ç̇midata : ʒmata : onisime : da solomonisi : epesis mtavarep̈isḳop̈ostay* ‘of Saint Ananias the Persian and of the Saints, the brothers Onesimus and Solomonus, the archbishops of Ephesus’. Note that the word for ‘saint’ in its singular and plural forms appears abbreviated here, with a similar overbar marking the suspension (*ç̃isa / ç̃ta*), as do many other common words in both the rubric and the main text. What follows in the fourth line of the heading, are elements of prayers (*upalo čueno* ‘our Lord’, again abbreviated: *õo čno*; and *šeiç̇qalen* ‘have mercy’, written *š̃n* with the first character in red and the second together with the abbreviation mark, in black); between them we find the indication of a ‘mode’ to be used in singing (*: q̃y :* = *qmay ã*, lit. ‘tone (or voice) 1’, with the noun written in black), and, as the first textual elements written in minuscules, the (abbreviated) *incipit* of the master hymn (*heirmos*) sung in that ‘mode’ (*s̈ṭ̃q̈y dau:* = *s̈iṭq̈uay dausabamoy*, ‘the boundless word’).

The main text block then consists of hymns of praise addressed to the commemorated saints, with the initial letters of the individual strophes extending into the margin, as majuscules; the first initial is in red, the others in black. The red dots (or combinations of red and black dots) denote boundaries between individual verses while the end of strophes is indicated by more complex arrangements of punctuation marks (⁖, ÷, and the like, in black). The most complex arrangements of dots, quincunxes (⁘) in black with a red cross overlaid, are found in the left margin, encircling the long-shaft cross in purple; as a matter of fact, the latter is likely to represent a character rather than the cross, namely the Georgian majuscule letter *k* (ⴊ) standing for ‘Christ’, *krisṭe*, or even its Greek equivalent, the *Chi-Rho* symbol, adapted in shape to the Georgian *k*.

With up-to-date computer systems and text processing software, it may well be attempted to reproduce the contents of a manuscript page of this complexity as it is, both on the screen and on a (colour) printer; Table 0.2.12 shows to what extent the ‘WYSIWYG’ principle (‘what you see is what you get’) can be achieved having appropriate fonts at hand. It must be stressed here, however, that some of the characters implied are not yet represented in Unicode (as of January 2014) so that the encoding remains arbitrary to a certain extent. This is true, for example, for the peculiarly shaped *k* symbol (with a loop to the right at its top) standing for *krisṭe*, which is replaced by a mere *k*-letter here (Unicode does provide a code point for the *Chi-Rho* symbol, U+2627, which might as well have been used). It is also true for the combinations of a quincunx with an overlaid cross (the former does have a code-point, U+2059, but the latter has none);

Table 0.2.12 Near-to-facsimile rendering of MS Vienna, ÖNB, Cod.Vind.georg. 2, front fly-leaf (excerpt)



what is more, the co-occurrence of two colours within the combinations makes it impossible to encode them as ‘precomposed’ characters. A more important deficiency of the Unicode standard of today is the lack of code-points for the different types of overbars appearing in numerical notations and abbreviations (over one character, over two characters, etc.) in mediaeval manuscripts, not only of Georgian provenance.

It must be stated off-hand that such a near-to-facsimile representation of the contents of a manuscript has only a very limited use as it can only be deployed as part of a ‘diplomatic’ edition (see Ch. 3 § 3.11 for this type of editions). For most other purposes, the ‘surface-oriented’ aim to reproduce the visual appearance of a given manuscript page will be deemed subordinate to a consistent registration of the meaningful elements contained in the texts and their functions. This is true, first of all, for the preparation of ‘critical’ editions that are based on the collation of several manuscripts. In this process, described in more detail in Ch. 3 of the present handbook, one would typically ignore the distinction of majuscules and minuscules as well as the different colours and sizes as appearing in our example. Words written across line breaks (with or without hyphenation marks, as usual in Georgian manuscripts) would be re-joined; in addition, one would resolve the abbreviations and, possibly, also the values of numerically used letters. Depending on the specific conventions of the individual scholarly traditions, one might further adapt the system of punctuation appearing in the manuscript with that used in modern orthography, including the corresponding division into sentences (or, in the case of metrical texts, verses) and the use of a modern script. For the fly-leaf of the Vienna codex, we should thus arrive at a rendering like that displayed in Table 0.2.13a. For the purpose of illustration, the Table contains the same text in both the modern Georgian *mxedruli* script and in a Roman transcription; note that the Old Georgian digraph ⴱⴳ (lit. *ow*), which represents the plain vowel *u*, is rendered by the single letter უ = *u* here as usual in modern Georgian editions.

The rendering thus achieved consists only of the most basic elements of textual contents, namely words (separated by spaces), clauses and sentences (separated by punctuation marks), and paragraphs (separated by hard line breaks). A ‘plain text’ of this type can indeed be used for several purposes, as the basis for a ‘critical’ text edition to be produced, as the basis of collation with other witnesses (automatic or manual, see Ch. 3 § 2.2), or as the basis for (automatic) indexation (Ch. 3 § 2.6.5). For the latter purpose, however, the ‘annotation’ of some more information will be required, depending on what kinds of indexes are to be generated. For a mere word index that ignores the affinity of a given (inflectional) word form to the corresponding lexicon entry (the lemma), it will still be necessary to apply a system of reference to the individual units of the text, that is chapters, paragraphs, sentences and the like if the indexation is meant to refer to its ‘internal’ structure, or production units, folia / pages, columns, lines and so on if it is meant to refer to its ‘external’ representation in a given manuscript—without such a referencing system, the index would be a mere assemblage of all word forms occurring, which would be rather worthless, especially if the text has a considerable length. The establishment and application of a consistent referencing system may also be helpful for later comparison of a given text with parallel sources. A good example for this is the referencing system used for Biblical texts today, which consists of the indication of a given book, chapter, and verse, and which has substituted older systems such as that of the Ammonian section numbers. In an ideal case, the different systems of reference relevant to a given text should be combined with

Table 0.2.13 Rendering of Vienna, ÖNB, Cod.Vind.georg. 2, f. 1a (excerpt)

(a) Plain text rendering

თთუესა დეკემბერსა 1.	*ttuesa deḳembersa 1.*
წმიდისა ანანიასი სპარსისაჲ და წმიდათა ძმათა ონისიმე	*ċmidisa ananiasi sṗarsisay da ċmidata*
და სოლომონისი ეფესის მთავარეპისკოპოსთაჲ.	*ʒmata onisime da solomonisi epesis*
	mtavareṗisḳoṗostay.
უფალო ჩუენო! შეიწყალენ!	*upalo čueno! šeiċqalen!*
ქმაჲ 1: სიტყუაჲ დაუსაბამოჲ...	*qmay 1: siṭquay dausabamoy...*
შესხმითა სულიერითა, ანანიას ერნო, შევამკუდეთ	*šesxmita sulierita, ananias erno, ševamḳudet ḳovelni*
ყოველნი წმიდასა მოწამესა, რომელმან დათრგუნა ძალი	*ċmidasa moċamesa, romelman datrguna ʒali igi*
იგი უჩინოჲსა მის მტერისა და გჯრგჯნ-შემოსილი	*učinoysa mis mṭerisa da gwrgwn-šemosili ixarebs*
იხარებს ზეცას კრებულსა თანა უჯორცოთასა.	*zecas ḳrebulsa tana uqorcotasa.*
ურიცხუნი განსაკითხავნი...	*uricxuni gansaḳitxavni...*

(b) Overlapping hierarchies (non-compliant)

```
<line n='5'><hymn n='1'><strophe n='1'><verse
n='1'>šesxmita sulierita, ananias erno,</verse> <verse
n='2'>šev</line><line n='6'>amḳudet ḳovelni ċmidasa
moċamesa,</verse> <verse n='3'>romelman datrg</
line><line n='7'>una ...</verse>
```

(c) Overlapping hierarchies (compliant)

```
<line n='5' /><hymn n='1'><strophe n='1'><verse
n='1'>šesxmita sulierita, ananias erno,</verse> <verse
n='2'>šev<line n='6' />amḳudet ḳovelni ċmidasa
moċamesa,</verse> <verse n='3'>romelman datrg<line
n='7' />una ...</verse>
```

each other as in the online edition of one of the oldest Georgian codices, the so-called 'Sinai Lectionary' of the Universitätsbibliothek Graz (Austria), provided by the TITUS project (Graz, UBG, 2058/1; Gippert et al. 2007b), which provides the references both to the position in the manuscript ('Manuscript page' and 'line') and to that of the Gospel passage concerned ('Book', 'Chapter', 'Verse') side by side (see fig. 0.2.3). In addition, the online text contrasts the 'diplomatic' rendering of the manuscript text (in Old Georgian majuscules) with a transcript into 'modern' style (*mxedruli*). The index produced on this basis is incorporated in a search engine which can be accessed, for example, by clicking upon a word form (in *mxedruli*), which will yield a list of all occurrences of the given word form within the same text, with clear indication of their position (see <http://titus.uni-frankfurt.de/texte/textex.htm> for a description of the applicable methods of use of the TITUS search engine, and fig. 0.2.4 for the output of the query for Georgian *ċigni* 'book, epistle, letter').

More sophisticated types of annotations must be applied if an index is to subsume word forms under their respective lemmas and if it is meant to differentiate common nouns from several types of proper names (personal names, toponyms, ethnonyms etc.), as usual in modern text editions. In this case, the word forms in question must be 'marked up' in a special way, with the corresponding information being added in an underlying structure. This is the approach taken by the 'Text Encoding Initiative' (TEI), a 'consortium which collectively develops and maintains a standard for the representation of texts in digital form' (see <http://www.tei-c.org>) and which comprises, among others, a 'Special Interest Group' concerning manuscripts (see <http://www.tei-c.org/Activities/SIG/Manuscript/>). The foundation of the TEI approach, outlined in extensive 'Guidelines for Electronic Text Encoding and Interchange' (present issue: 'P5'; <http://www.tei-c.org/release/doc/tei-p5-doc/en/html>), is the application of the so-called 'eXtensible Markup Language' (XML), an extremely flexible markup system developed by the 'World Wide Web Consortium' (W3C; <http://www. W3.org/XML/>) since the 1990s in extension of former standards such as SGML ('Standard Generalized Markup Language') and HTML ('Hypertext Markup Language', the markup system used predominantly in web pages to this day). The basic structural element of these markup languages consists of so-called 'tags', i.e. information units stored, in angle brackets, either on both sides of a text element to be marked up ('start-tag' and 'end-tag') or as independent entries ('empty-element tag'); these tags will usually not be rendered as such on the screen or in print but serve the purpose of controlling the output 'from behind'. To mark, for example, that a given word in a text is meant to be output in bold characters in an HTML-based web page, it has to be enclosed in two corresponding tags, which are and respectively, denoting the beginning and the end of the bold-faced area. With an empty-element tag, one can add the information that there is a line-break at a given position; the corresponding HTML tag is
. In contrast to this, XML exhibits two differences. First, empty-element tags must here be terminated by a slash within the brackets (
), thus distinguishing them from start-tags, which have no slashes. Second, and this is the major advantage of XML, the tags to be used can be chosen *ad libitum*, provided the choice is declared in either a 'Document Type Definition' (DTD)

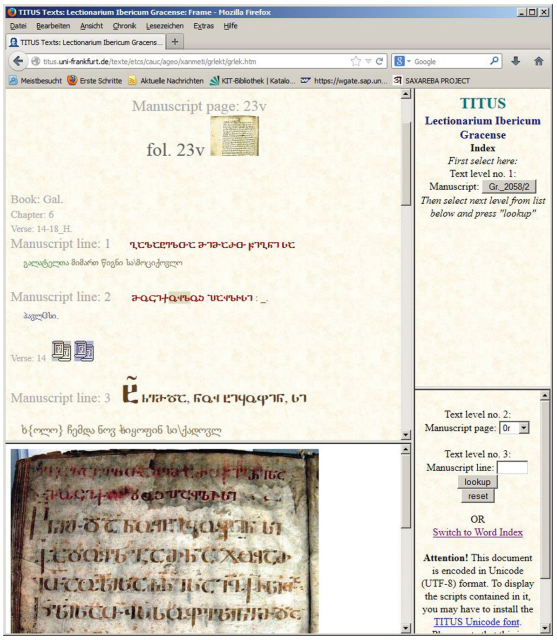


Fig. 0.2.3 Online edition of the Graz Sinai Lectionary

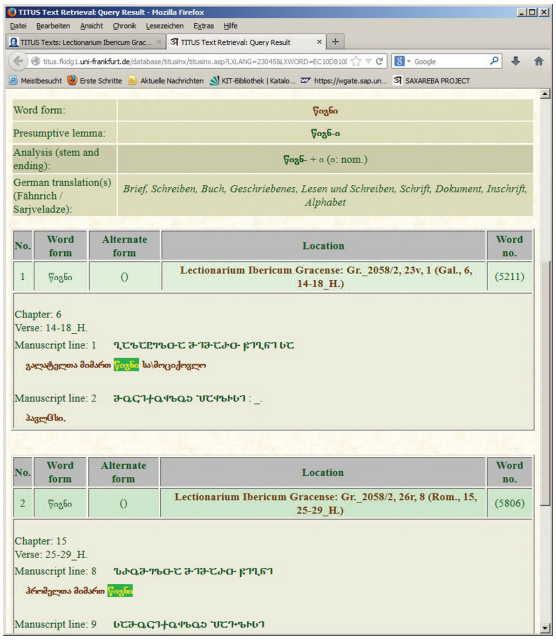


Fig. 0.2.4 Search engine output (*çigni* 'book')

or an 'XML Schema Definition' (XSD). This allows, for example, the use of a more explicit <bold> tag instead of <b>, or <line-break /> instead of
. Unlike the fixed set of tags acknowledged by the HTML standard, which was mostly addressed towards screen output and did not therefore contain many content-related tags, XML can thus be conceived to further distinguish several types of meaningful text elements such as indications of dates (in our Georgian menaion example, <date>*ttuesa deḳembersa* 1</date>), personal names (for example, <anthroponym>*ananiasi*</anthroponym>), ethnonyms (<ethnonym>*sṗarsisay*</ethnonym>), hymn incipits (<incipit>*siṭquay dausabamoy*</incipit>), or verses (<verse>*šesxmita sulierita, ananias erno*</verse>), with a view to a particular rendering in the output, to proper indexation, or to other purposes.

An even more powerful feature of the markup languages is the possibility of adding 'attributes' to the tags, consisting of a denominator and a value. These can be output-oriented as in the case of the HTML 'font' tag which can imply information as to the size, colour, and other features of the font the marked-up text is to be displayed in (in our manuscript, for example, *deḳembersa*). Beyond this, an appropriate XML tag may contain lexical, grammatical, or other content-related information (for example, <word lemma='deḳemberi' morph='dat_sg'>*deḳembersa*</word>). The flexibility of XML even allows for a combination of both types of information (<word lemma='deḳemberi' morph='dat_sg' fonttype='mrglovani' fontsize='12' fontcolour='red'>*deḳembersa*</word>). By the way, it is true that much 'markup' information that is linguistic can be added automatically, by applying so-called 'parsers' that analyse the given text on the basis of programmed grammatical rules and lexicons; however, in the field of oriental manuscript studies and the languages relevant to them, the development of tools for these purposes is not yet very much advanced.

Another important feature of XML is that taggings can further be nested, thus allowing, for example, to account for the change of the font colour in the abbreviated imperative form *šeiçqalen* 'have mercy' in our text, which might be tagged as <word expanded='šeiçqalen' lemma='šeçqaleba' morph='impv_aor'><chunk fontsize='14' fontcolor='red'>*š*</chunk><chunk fontsize='14' fontcolor='black'>*̃n*</chunk></word>. (As a matter of fact, several more sophisticated ways of annotating abbreviated word forms have been designed in the TEI-P5 guidelines.)

A peculiar problem of XML is that hierarchically organized taggings must not overlap in the sense that a start tag Y must not fall between a superior sequence of a start tag X and an end tag X if the end tag corresponding to Y does not (schematically: <X> … <Y> … </X> … </Y>). This is especially crucial for the parallel markup of different referencing systems ('internal' and 'external' references in the sense outlined above). If in our Georgian example, we wanted to mark up both the units of the text structure (for example, verses) and their distribution on the manuscript page, we should arrive at exactly this problem right from the second verse on; what is more, there are line breaks within words that would have to be accounted

for. Table 0.2.13b shows the resulting picture for the first three lines of the hymnal text, which would not be XML-compliant. A possible way out of this is the use of empty-entity tags for one of the overlapping hierarchical referencing systems; in Table 0.2.13c, it is the ('external') line referencing that is treated this way, with an XML-compliant result (note that font colours and other similar parameters are ignored here).

Taking all the features of XML together, it is conceivable that the contents of a manuscript can be electronically annotated with them in such a way that both different forms of editions ('diplomatic' and 'critical', printed and online) and several kinds of indexes can automatically be derived from the annotated text (cf. Ch. 3 § 3.1 for relevant considerations). For the former purpose, this presupposes the design and application of so-called 'Extensible Stylesheet Language Transformations' (XSLT), which can be used to transform XML documents into HTML web pages, plain text files, or 'Extensible Stylesheet Language Formatting Objects' (XSL-FO) which can subsequently be converted to PDF or other output formats. For indexation, one may still have to rely upon special tools that are conceived to extract the targeted information. The more scholars show interest in these kinds of tools and methods, the more it is likely that we shall have them at hand for usage in the foreseeable future.

2.1.3. Manuscript related databases and their structuring

In recent years, XML has gained more and more ground in yet another domain that is relevant for manuscript studies, namely the structuring of databases. If we leave indexes used for the search of words or word forms in textual contexts aside, the typical field of application for databases concerning manuscripts is cataloguing. More and more manuscript catalogues are being conceived and compiled electronically today, both as a basis for printed output and for the integration in online search engines, portals, and the like (see Ch. 4 § 6), and the question of how to structure them may therefore be crucial. As in other fields of application, XML-based structures are in competition with so-called 'relational' databases here, and the decision in favour of one or other of them may not be easy to take.

The main difference between the two types of database consists in the fact that XML yields more flexible structures than relational databases, which are characterized by a consistent setup of 'records', that is entries. Typically, a record in a relational database comprises a fixed set of 'fields' that are identical throughout the whole collection of data of the same structure. The interrelation of these elements can easily be visualized in form of a table, with the rows representing records and the columns, fields; see Table 0.2.14 for an arbitrary example that is derived from the description scheme developed for the 'Union Catalogue of Oriental Manuscripts in German Collections' of the Göttingen Academy of Sciences (see Ch. 4 § 6.1 for more details). It is true that such a scheme can be extremely helpful to ascertain that no item of information is overlooked and that the data are kept consistent, for example, in their orthographical representation, throughout the records; there is a clear disadvantage, however, in that it may be difficult, if not impossible, to deal with manuscripts of mixed content, written by different scribes and/or at different times and places, etc. In other words, as soon as we take codices into account that consist of several 'production units' (see Ch. 4 § 4 for the concept underlying this term), the given scheme may all too soon prove to be too rigid to be expedient.

If we conceive the same database in an XML structure, we may indeed 'spread' the scheme much more easily according to the peculiarities of our objects. The 'shelf number' may still be the governing information, but we may insert any number of 'production units' below it, each with its own record of data. In addition, there is no limit as to the amount of data to be stored within a given field, different from relational databases where this may lead to problems. Table 0.2.15 may give an idea of such an approach, building upon the arbitrary example introduced above.

It will be clear from this example that an XML database has a certain disadvantage, too. This is the amount of data that has to be stored and processed in a clear-text structure of this type. This may be un-

Table 0.2.14 Relational database structure used in cataloguing (example)

shelf number	material	state of preservation	pages	format	lines	writing style	decoration	scribe	date	origin	author	title
1	parch.	III	142	17 × 23	26	maj.	+	Io.Zos.	981	Sinai	anon.	Gospels
2	paper	II	255	16 × 24	29	min.	–	unknown	1231	Šaṭberdi	Mi.Mo.	Hymn.
3	parch	IV	183	18 × 23	25	maj.	+	Io.Xax.	X	Ṭao	anon.	Hagio.

Table 0.2.15 XML database structure used in cataloguing (example)

```
<shelfnumber n='1'>
  <productionunit n='1'>
    <material>parchment</material>
    <stateofpreservation>III</stateofpreservation>
    <pages>1r-126v</pages>
    <format>17 × 23</format>
    <lines>26</lines>
    <writingstyle>majuscules</writingstyle>
    <illumination n='1' page='3r'>Matthew</illumination>
    <illumination n='2' page='38r'>Mark</illumination>
    <illumination n='3' page='64v'>Luke</illumination>
    <illumination n='4' page='101r'>John</illumination>
    <scribe>Ioane Zosime</scribe>
    <date>981</date>
    <origin>St. Catherine's Monastery, Mt. Sinai</origin>
    <author>anonymous</author>
    <title>Gospels</title>
  </productionunit>
  <productionunit n='2'>
    <material>parchment</material>
    <stateofpreservation>III</stateofpreservation>
    <pages>127r-139v</pages>
    <format>17 × 22.5</format>
    <lines>28</lines>
    <writingstyle>majuscules</writingstyle>
    <scribe>Ioane Zosime</scribe>
    <date>981</date>
    <origin>St. Catherine's Monastery, Sinai</origin>
    <author>anonymous</author>
    <title>Lection index</title>
  </productionunit>
  <productionunit n='3'>
    <material>parchment</material>
    <stateofpreservation>IV</stateofpreservation>
    <pages>140r-142v</pages>
    <format>17 × 22.5</format>
    <lines>29</lines>
    <writingstyle>minuscules</writingstyle>
    <scribe>Ioane Zosime</scribe>
    <date>981</date>
    <origin>St. Catherine's Monastery, Sinai</origin>
    <author>Ioane Zosime</author>
    <title>Colophon</title>
  </productionunit>
</shelfnumber>
<shelfnumber n='2'>
  <productionunit n='1'>
    <material>paper</material>
    <stateofpreservation>II</stateofpreservation>
    <pages>1r-255v</pages>
    <format>16 × 24</format>
    <lines>29</lines>
    <writingstyle>minuscules</writingstyle>
    <scribe n='1'>unknown</scribe>
    <scribe n='2'>Giorgi</scribe>
    <date>1231</date>
    <origin>Šaṭberdi</origin>
    <author>Mikael Modreḳili</author>
    <title>Hymnary</title>
  </productionunit>
</shelfnumber>
...
```

problematic if the database is only meant to be the basis for printed or online output; for other purposes such as, for example, retrievability via hypercatalogues (see Ch. 4 § 6.2), relational databases may still be regarded as superior, given that they can be accessed much faster due to preindexation. However, with the steadily increasing storage capacity and processing speed of modern computers, this advantage may vanish soon.

2.1.4. Digital imaging

No field relevant to oriental manuscript studies has profited more from technological progress in the digital age than imaging. A clear witness to this is the fact that the amount of high-quality images of manuscripts that are available online has been increasing exponentially since the late 1990s, and many of us use such images every day without thinking too much about their structural properties. Nevertheless, it may be worthwhile here to summarize a few basics concerning the processes involved.

No matter what quality is to be achieved, digital imaging presupposes the dissolution of the visual appearance of a given object into a bulk of tiny dots, so-called pixels, each of them characterized by a certain degree of light intensity of different colour components, mostly red, green, and blue, exposed either individually or in groups (stacks). The number of picture cells (pixels) available on the camera sensor is the basis for the calculation of the data a digital image comprises, usually called its 'resolution': while by the end of the twentieth century, an amount of two megapixels (1,600 × 1,200 pixels, with an aspect ratio of 4:3) was still beyond reach, cameras with a resolution of 50 megapixels (8,176 × 6,132 pixels with the same ratio) are no longer exceptional today. With such a resolution, even an extremely large manuscript page of 82 × 61 cm could be photographed and reproduced in printed form without any visible loss of information, the resolution still yielding 10 pixels per millimetre in printing. For the complete rendering of the same page on a computer screen, much lower resolutions would be sufficient, given that a normal screen resolution of 1280 × 960 pixels equals to no more than 1.23 megapixels; however, the great advantage of large-resolution digital images is that they can be enlarged in screen output so that individual sectors of the manuscript page can be displayed in even much larger size than that of the original.

The calculation by pixels (or dots) per centimetre (or per inch, differentiated by a factor of 2.54) may be misleading, however. In the early years of manuscripts digitization, when the resolution of digital cam-

eras was not yet sufficient for this purpose, attempts were made to achieve the same goal by applying optical scanners with much lower resolutions; for example, a flatbed scanner with a surface of 21 × 29.7 cm (the measure of A4 paper) and a resolution of 600 dots per inch (dpi; the metrical equivalent would be 236 dots per centimetre) yielded a digital image of (4960 × 7015 =) 34.8 megapixels, and even with 300 dpi the image still had (2480 × 3057 =) 8.7 megapixels. However, the application of flatbed scanners for the digitization of manuscripts was not always possible due to conservation concerns, either because of the extreme light exposure those scanners work with or because of the threat of damaging the binding of the codices etc. Therefore, an intermediate solution was sought in the application of a hybrid approach which made use of traditional (film) photography by producing colour slides as the basis for digitization; this approach was, for example, applied in one of the earliest projects aiming at an online edition of manuscripts comprising colour images of the originals, namely the project concerning the Tocharian manuscripts of the Berlin Turfan collection, which have been published on the TITUS server since 1999 (see <http://titus.fkidg1.uni-frankfurt.de/texte/tocharic/tht.htm>). The resolution that could be achieved on this basis in the late 1990s was 2700 dpi, a value seeming much higher indeed than the 600 dpi of a flatbed scanner; however, we must consider that the surface of the underlying colour slides was much smaller than that of any manuscript page and that the scanner resolution is always relative to the size of the scanned object: when a colour slide of 24 × 36 mm containing an image of an A4-sized manuscript page was scanned at the resolution of 2700 dpi, the resulting image comprised (2551 × 3827 =) 9.7 megapixels, which was not much more than the resolution of a 300 dpi scan of the same page on an A4 flat-bed scanner or a digital image of it with a resolution of 10 megapixels (Table 0.2.16 lists some noteworthy figures concerning the digitization of an A4-sized manuscript page). Still, the production of colour slides had a big advantage, given that they can be used as a secondary ('analog') storage medium in order to preserve the contents of a large amount of manuscripts and that they remain available for scanning with higher-resolution scanners for a long time, with no need to touch (and contaminate) the original documents. It must be underlined though that all this depends on the quality of the film used and that only a few colour slide films have proven to sustain the quality of the images they contain over a longer period of time.

The same holds—and even more so—for the digitization of microfilms, an approach that has been undertaken with great effort until the present day (for example, the digital collections of manuscripts at the Bayerische Staatsbibliothek, Munich, are partly based on microfilms 'in a bitonal or grey-scale quality' instead of 'full colour copies' of the original manuscripts; see <http://www.digitale-sammlungen.de/index.html?c=sammlungen&kategorie_sammlung=1&l=en>). This may be acceptable in cases where the original manuscripts have been lost or are no longer or not easily accessible for other reasons, as in the case of the microfilms of the manuscripts of St Catherine's Monastery on Mount Sinai which were produced in the 1940s on behalf of the Library of Congress and parts of which have now been digitized for online retrieval (see, for example, the collection 'Microfilms des manuscrits géorgiens du Mt Sinai' provided by the Université Catholique de Louvain, Belgium, <http://www.e-corpus.org/eng/notices/96559-Microfilms-des-manuscrits-g%C3%A9-orgiens-du-Mt-Sinai.html>). In other cases, however, the quality of microfilms, especially those produced during extensive microfilming campaigns as in the case of the Sinai manuscripts, is hardly sufficient to meet the requirements of in-depth manuscript studies. This is all the more true since the microfilms used in such campaigns were usually monochrome, thus obscuring information on the use of different (coloured) inks, which may be crucial as a text structuring element in many cases (see above). In any digitization project, the question of whether and to what extent microfilms may be a usable basis should therefore be pondered seriously. The production of new digital images directly from the original manuscripts will nearly always yield much better results today (see also Ch. 5 § 7 for a detailed treatment of the processes involved).

In the recent past, special methods of digital imaging have gained importance in oriental manuscript studies, especially in the analysis of palimpsests. Based on the fact that parchment as the typical support material of palimpsests fluoresces in ultraviolet (UV) light (see General introduction § 2.3), it was mostly UV photography that was used until the end of the twentieth century to enhance the contrast between the parchment surface and the ink of the underwriting, with more or less satisfying results. By the beginning of the twenty-first century, UV photography has been superseded by so-called 'multispectral imaging', a process that builds upon the production of several images that are restricted to a certain wavelength of the visible and the invisible light (ultraviolet and infrared), and the digital comparison of these images. The main principle of multispectral imaging consists in the fact that the resonance of any object differs with

Table 0.2.16 Digitizing a manuscript page of A4 size

A4-page	11,69 × 8,27 inch	29,7 × 21,0 cm
colour slide / microfilm image	1,42 × 0,94 inch	3,6 × 2,4 cm
Microfilm / slide scanner, 1200 dpi	1704 × 1132 pixels	2 megapixels
Flatbed scanner, 300 dpi	3507 × 2480 pixels	8.7 megapixels
Microfilm / slide scanner, 2700 dpi	3834 × 2538 pixels	9.7 megapixels
Digital camera, 12 megapixels	4200 × 2800 pixels	11.7 megapixels
Microfilm / slide scanner, 4000 dpi	5680 × 3760 pixels	21.35 megapixels
Flatbed scanner, 600 dpi	7014 × 4960 pixels	34.8 megapixels
Flatbed scanner, 1200 dpi	14028 × 9920 pixels	139.2 megapixels

respect to different wavelengths, depending on the consistence of its colour. By applying a photographing method that is restricted to a certain range of the spectrum, a specific resonance may be retained or suppressed. In the case of palimpsest manuscripts, the effect that can be gained from this predisposition depends on three factors: the colour resonance of the upper script, that of the lower script, and that of the background, i.e. the parchment surface. One might expect that the first two are the most decisive factors in this constellation, as in many cases it will be desirable to 'enhance' the lower script in contrast to the upper script covering it. This, however, is not always possible in parchment palimpsests of oriental provenance as both the lower and the upper scripts were usually written with the same type of inks, which results in similar resonances. Thus the application of multispectral imaging must concentrate upon two aims, a) increasing the contrast between the (erased) lower script and the background, and b) exploiting the difference of several images showing the same object to reduce the preponderance of the upper script. Normally, a set of three images (one in the UV or violet range, at a wavelength of less than 440 nm; one in the yellow or green range, at a wavelength of between 500 and 600 nm, and one in the red or near-infrared range, at a wavelength of above 700 nm) will be sufficient for this purpose. Several projects concerning oriental palimpsests have successfully adapted multispectral imaging since 2002 (see General introduction § 2.4), and the methods and facilities implied are steadily developing.

References

Gippert et al. 2007a. Web sources: Gippert et al. 2007b; <http://www.digitale-sammlungen.de/>, last access October 2014; <http://www.e-corpus.org>, last access October 2014; <http://www.tei-c.org/>, last access October 2014; <http://www. W3.org/XML/>, last access October 2014; <http://titus.fkidg1.uni-frankfurt.de>, last access October 2014