Volume 27 - Number 2 - 2012 - ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Altüberlieferte Sprachen als Gegenstand der Texttechnologie

Ancient Languages as the Object of Text Technology

Herausgegeben von /*Edited by*
Armin Hoenen und Thomas Jügel

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 27 – 2012 – Heft 2
Herausgeber	Armin Hoenen und Thomas Jügel
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Altüberlieferte Sprachen als Gegenstand der Texttechnologie

Ancient Languages as the Object of Text Technology

Guest editors: Armin Hoenen and Thomas Jügel

Text technology predominantly focuses on modern languages. Most of these languages supply text technology with a significant amount of digitised texts. Constantly growing internet resources such as various Wikis produce new analysable data, which serve as input and testing grounds for statistical models, rule-based analyses and various other research tools. While the number of available tokens might easily reach billions, historians usually deal with a fixed set of data and can deem themselves lucky if their texts are complete. If, then, the corpus size reaches to several thousands of words, it is even better.

The aim of this volume is to represent several projects on historical corpora: Corpus Avesticum (Old Iranian), Mercurius Corpus of Early New High German, Referenzkorpus Altdeutsch (Old German Reference Corpus), Old Lithuanian Reference Corpus (SLIEKKAS). As such, the focus is on the humanities' perspective, so to say, on the user's side of text technological tools.

Historical corpora can have several drawbacks, and their evaluation depends heavily on the expertise of specialists. Such corpora are inherently limited, texts might be incomplete, it is likely that we only have an incomplete knowledge of the grammatical system, and there remain uncertainties in lexical meaning. Language change cannot always be clearly located in space and time, so that ambigue phrases in ancient texts could allow for two different interpretations: an old and a new one, or as aptly put by PAULY et al. (this volume): "die Ausgangs- oder die Zielstruktur". Gaps in transmission (be it loss of material, or be it loss of text when transmitted orally), divergent variants due to the copying process, and the uncertainty of place, time, and author of texts all make the identification of originals a difficult task. Hence, the digitisation and computational evaluation of historical linguistic corpora faces problems that can issue challenges to text technology. It is obvious that techniques like machine learning suffer from lack of training data—but there is more to text technology.

With a good annotation for a corpus at hand (the building of which is a labour intensive task), more sophisticated analyses can be conducted by means of technological applications, such as a search tailored for the historian or linguist. Combined queries for word forms allow the investigation of grammatical rules: Which mood follows specific conjunctions? Which case is governed by prepositions? Can prepositions be used as postpositions as well? Which kind of word order is possible? Furthermore, an analysis of co-occurrences of a term under consideration in texts that belong to different eras can reveal semantic changes. Repetitions can reveal the original structure of a text (e.g., of a ceremony), if, e.g., these have been obscured by a modern division into chapters. Detecting intertextual dependencies enables us to trace the path a text took through time and languages, i.e., to trace its reception.

Annotation is the very starting point of historical text technology and, thus, it is a main focus of the present volume. The articles of MITTMANN and LINDE discuss automatic pre-annotation of glossaries of Old German and the problems that occur when defining tags for information from printed media. PAULY et al. examine the representation of ambiguous and discontinuous phrases in Early New High German under the scope of language development. Similar in method is the project on annotating Old Lithuanian texts by GELUMBECKAITĖ et al.

Another complex contains articles representing issues of the Old Iranian language Avestan. GIPPERT gives an overview of the encoding strategies of the complex Avestan writing system. JÜGEL considers the problems concerning the automatic generation of stemmata for Avestan manuscripts. By means of this volume, we hope to bring the disciplines of humanities and text technology closer to one another and to support the exchange of information between these fields of study.

At long last, we would like to thank Dr. Timothy Price, for he not only checked the English articles for spelling, flow, and style, but also gave many helpful comments that helped to improve the articles enormously. We would also like to express our gratitude to Dr. Lothar Lemnitzer and the *Gesellschaft für Sprachtechnologie und Computerlinguistik* for accepting this volume to be published in the JLCL series. Last but not least, our appreciation goes to the many reviewers who so generously spent their time in helping us to improve our articles.

The Editors
Frankfurt am Main, 2012

Content

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]

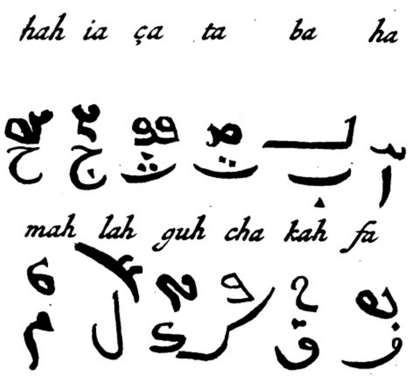


Fig. 1: CHARDIN's alphabet list (extract)

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]

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 melius eluceſcant,) ex adverso apposui nomina Medica quibus tales hodiè appellantur, seq. modo; simul cum quibusdam aliis:

Lingua Medica, qua Persæ audit.			*Lingua Tablavi, qua verè Persica.*
پارسا *Pârsâ*	*Devotus*,	*Religiosus*	*Catûzi*
سپاهي *Spâhi*		*Militaris*	*Neisâri*
دهکان seu دهقان *Dihcân*		*Agricola*	*Nasûdi*
دانا *Dânâ*		*Doctus*	*Ahanûchashi*
خدا *Chodâ*		*Deus*	*Yezdân*
خدا *Chodâ*	*Deus*	Ὡρομάσδης	*Hormùzd*
فرشته *Phríshta*	*Angelus*	*Amshâspand*	
دیو *Dîv*	*Diabolus*	Ἀρειμάνιος	*Aharîman*
تشنگي *Téshnaghi*		*Sitis*	*Tushnamâr*
گرسنگي *Gurísnaghi*		*Fames*	*Gushnamâr*
شستگي *Shúshtaghi*		*Lotio*	*Pâdyâvand*
خاموش *Châmûsh*		*Silentium*	*Bâgj* seu *Bâzh*

Primi Ordinis کاتوزیان *Catuzæos* constituit in Montibus & Speluncis vel habitare vel multùm versari, ibique cultui divino & meditationi atque doctrinæ cœlestis acquisitioni vacare. Nam hoc erat longè ante *Zerdùshst*, sc. tempore Sabaismi, quo Deum colebant (& simul Planetis adolebant) in Montibus, vel ut Strabo ait ἴδρυον ἐν ὑψηλῷ τόπῳ. Secundi Ordinis نیسارزیان *Neisaræos* voluit seipsos dedicare τῇ سپاهگري *Spâhigheri* seu *Militiæ*. Tertii Ordinis نسودیان *Nasûdæos* voluit Agriculturæ operam dare. Quarti Ordinis آهنوخشیان *Ahanûchashæis* præcepit omnis generis Literaturæ operam dare, sc. bonas Literas colere & studium in iis collocare. Istam etiam Quadruplicem po-

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

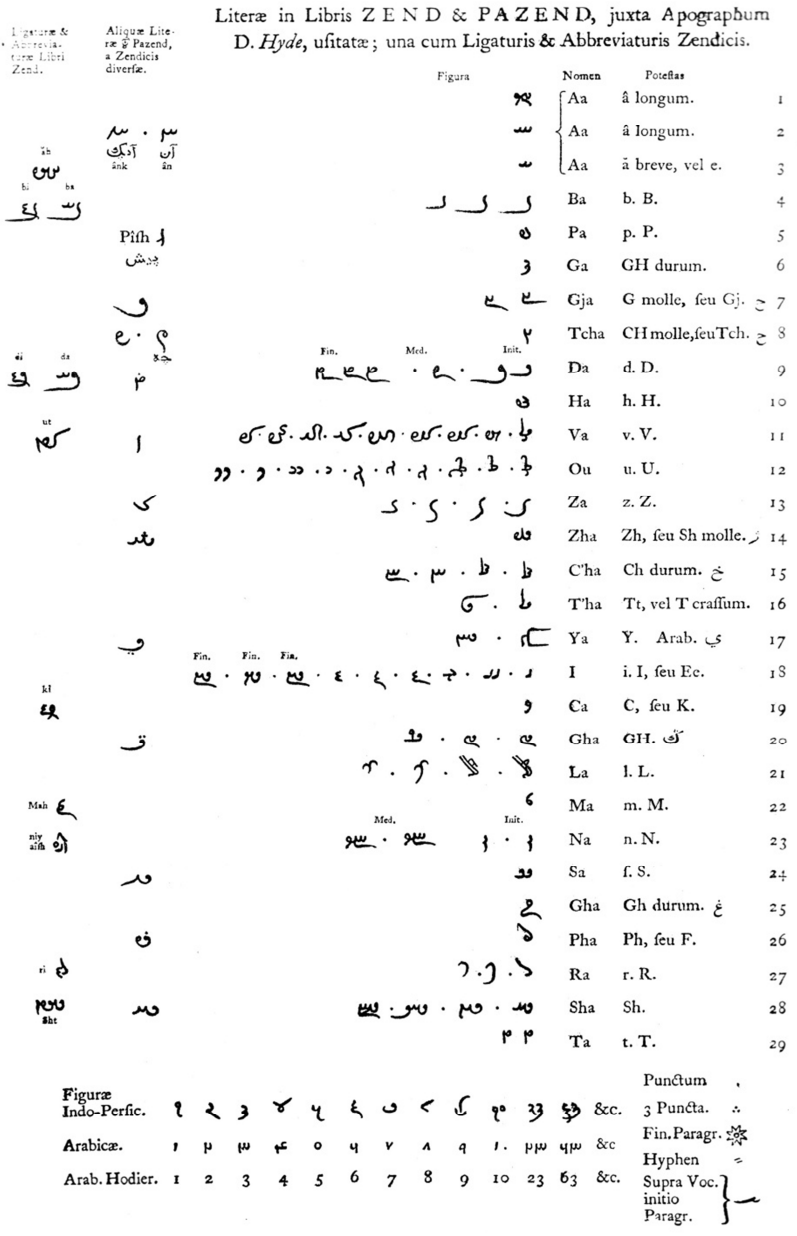


Fig. 3: Alphabet list in HYDE, Historia, 1760.

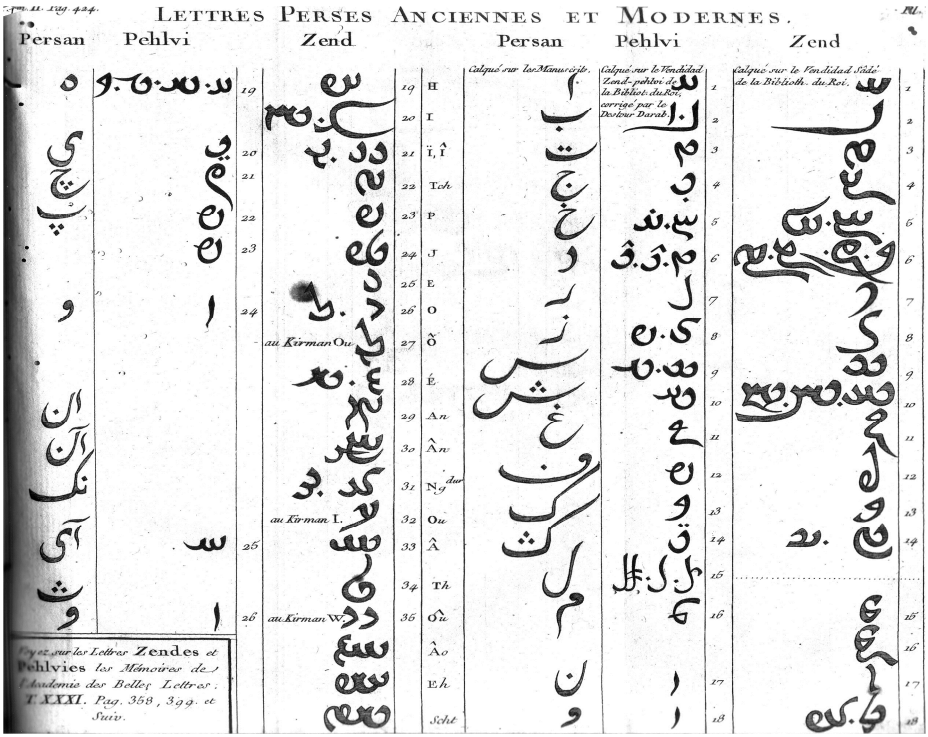
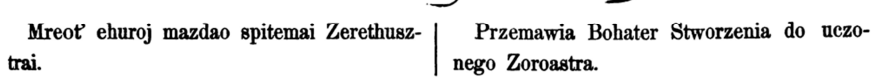


Fig. 4: Alphabet list in ANQUETIL-DUPERRON, Zend-Avesta



Mreot' ehuroj mazdao spitemai Zerethusztrai.

Przemawia Bohater Stworzenia do uczonego Zoroastra.

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

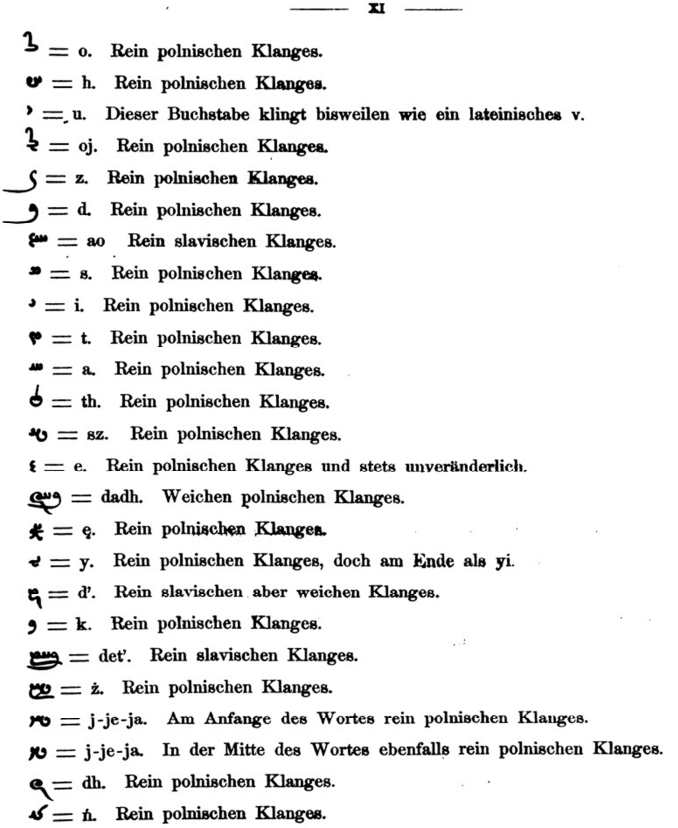
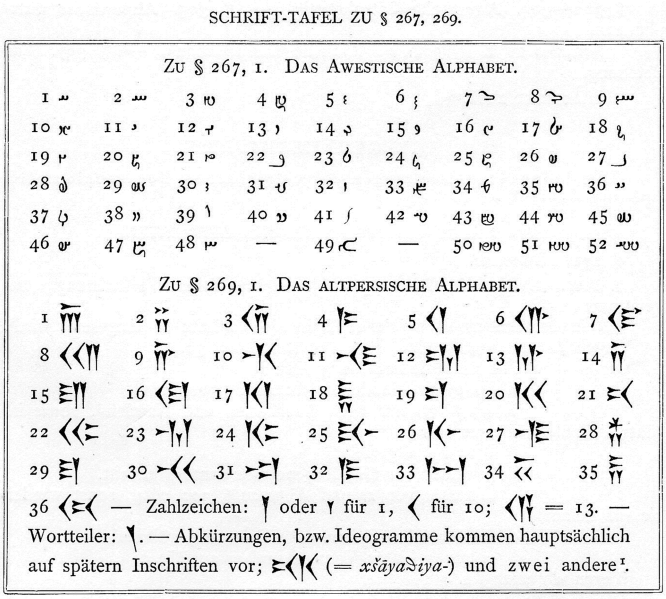


Fig. 6: PIETRASZEWSKI's transcription

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.

1 𐬀 2 𐬁 3 𐬂 4 𐬃 5 𐬄 6 𐬅 7 𐬆 8 𐬇 9 𐬈
10 𐬉 11 𐬊 12 𐬋 13 𐬌 14 𐬍 15 𐬎 16 𐬏 17 𐬐 18 𐬑
19 𐬒 20 𐬓 21 𐬔 22 𐬕 23 𐬖 24 𐬗 25 𐬘 26 𐬙 27 𐬚
28 𐬛 29 𐬜 30 𐬝 31 𐬞 32 𐬟 33 𐬠 34 𐬡 35 𐬢 36 𐬣
37 𐬤 38 𐬥 39 𐬦 40 𐬧 41 𐬨 42 𐬩 43 𐬪 44 𐬫 45 𐬬
46 𐬭 47 𐬮 48 𐬯 — 49 — 50 𐬰𐬀 51 𐬰𐬁 52 𐬰𐬂

ZU § 269, 1. DAS ALTPERSISCHE ALPHABET.

1 𐎠 2 𐎡 3 𐎢 4 𐎣 5 𐎤 6 𐎥 7 𐎦
8 𐎧 9 𐎨 10 𐎩 11 𐎪 12 𐎫 13 𐎬 14 𐎭
15 𐎮 16 𐎯 17 𐎰 18 𐎱 19 𐎲 20 𐎳 21 𐎴
22 𐎵 23 𐎶 24 𐎷 25 𐎸 26 𐎹 27 𐎺 28 𐎻
29 𐎼 30 𐎽 31 𐎾 32 𐎿 33 𐏀 34 𐏁 35 𐏂
36 𐏃 — Zahlzeichen: 𐏑 oder 𐏒 für 1, 𐏓 für 10; 𐏔 = 13. — Wortteiler: 𐏐. — Abkürzungen, bzw. Ideogramme kommen hauptsächlich auf spätern Inschriften vor; 𐏃 (= *xšāyaϑiya-*) und zwei andere[1].

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

EINLEITUNG: DAS SCHRIFTWESEN.

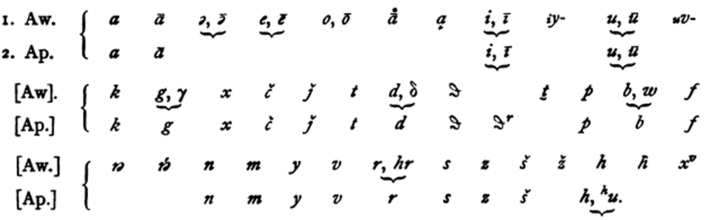
I. DAS AWESTISCHE SCHRIFTWESEN.

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 *å* 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 *ṯ* 26 *ṗ* 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 *ẖ* 48 x^v; 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*	*å*	*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*		*ṯ*	*ṗ*	*b, w*	*f*		
[Ap.]	*k*	*g*	*x*	*č*	*ǰ*	*t*	*d*	*ϑ*	$ϑ^r$		*ṗ*	*b*	*f*		
[Aw.]	*ŋ*	*ŋ́*	*n*	*m*	*y*	*v*	*r, hr*	*s*	*z*	*š*	*ž*	*h*	*ẖ*	x^v	
[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]

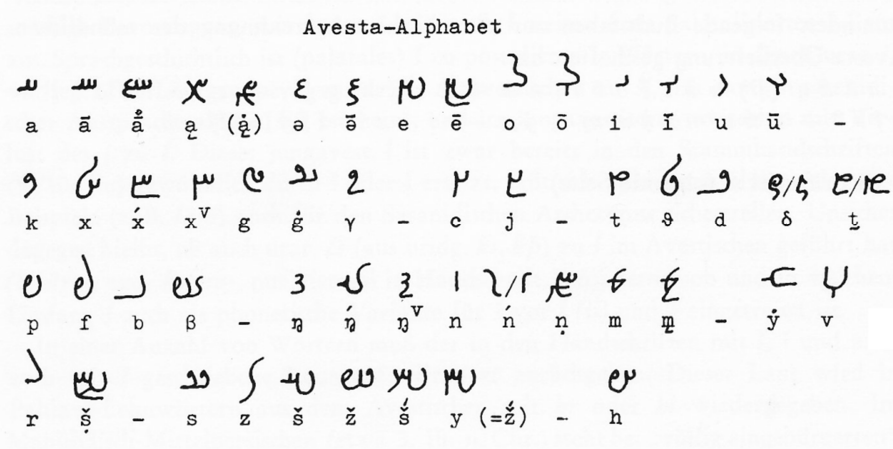


Fig. 10: HOFFMANN's 'Zeicheninventar'

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*	*å̄*	𐬃
𐬄	29	an	—	10	*ą*	*ą*	𐬄
𐬅	—	—	ę	—	—	*ą̇*	𐬅
𐬆 [19]	—	—	—	—	—		𐬆 [19]
𐬇	28	e	e	5	*ə*	*ə*	𐬇
𐬈	—	—	—	6	*ə̄*	*ə̄*	𐬈
𐬉	28	e	j-je-ja	3	*e*	*e*	𐬉
𐬊	—	—	je	4	*ē*	*ē*	𐬊
𐬋	26	o	o	7	*o*	*o*	𐬋

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

Original	Anqu.-No.	Anqu.	Pietr.	Bthl.-No.	Bthl.	Hoffmann	Original
𐬬	20	i	—	49	*y*	*ẏ*	𐬬
𐬭			j-je-ja	35		*y*	𐬭
𐬮	21	ï, î	y	36		*ii*	𐬮[20]
𐬯	18	v	—	37	*v*	*v*	𐬯
𐬰[19]	—	—	—	—			𐬰[19]
𐬱	35	oˆu	w	38		*uu*	𐬱[20]
𐬲	7	r	r	39	*r*	*r*	𐬲
𐬳	—	(l)	(l)	—	—	—	𐬳
𐬴	9	s	s	40	*s*	*s*	𐬴
𐬵	8	z	z	41	*z*	*z*	𐬵
𐬶	10	sch	sz	42	š	*š*	𐬶
𐬷			—	44		*ś*	𐬷
𐬸			ż	43		*ṣ*	𐬸
𐬹	24	j	ż	45	ž	ž	𐬹
𐬺	19	h	h	46	*h*	*h*	𐬺
𐬻[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 $ = *š*, 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 'jasrayā́ sūrmyā̀ yaviṣṭha / tvā́ṃ śáśvanta úpa yanti vā́jāḥ*

Table 3: Encoding of the Texas Ṛgveda (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	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
			0												1					
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	{	|	}	~	≈	∵	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	ǵ	ł	ű	à	ě	ì	ı
180	ů	ą	ǟ	ǽ	x^{v}	ž	ŋ^{u}	ṛ	ĭ	ī	ŭ	ą	ę	į	ǫ	ų	į	ų	ə	ə̄
200	ə̗	ą̃	ą́	ę̃	ę́	ẽ	é	į̃	į́	ų̃	ú	ũ	ỹ	ý	β	ƀ	č	ḍ	đ	δ
220	ǵ	ġ	g̱	γ	ḫ	ß	ḥ	ʰu	ḱ	ḷ	ĺ	Ī	Ĩ	ṃ	m̃	m̆	ṃ	ṅ	ń	ṇ
240	ṇ	ṛ	ŕ	r̄	ŕ	r̃	ś	ṣ	š	ś	ṣ	ṯ	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	Ǎ	1,162	Ő
1,24	ı	1,91	ǎ	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	Đ	1,178	Ş
1,40	É	1,107	ě	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	Ģ	1,186	Ŧ
1,48	Í	1,115	ģ	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	Ĝ	1,190	Ű
1,52	Ï	1,119	ĝ	1,191	ű
1,53	ï	1,120	Ġ	1,192	Ū
1,54	Ì	1,121	ġ	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	Ĭ	1,208	Ż
1,70	Ü	1,137	ĭ	1,209	ż
1,71	ü	1,138	IJ	1,210	Ð
1,72	Ù	1,139	ij	1,211	ŋ
1,73	ù	1,140	Ĵ	1,212	Đ
1,74	Ÿ	1,141	ĵ	1,213	đ
1,75	ÿ	1,142	Ķ	1,214	L̄
1,76	Ã	1,143	ķ	1,215	l̄
1,77	ã	1,144	ĸ	1,216	N̄
1,78	Ð	1,145	Ĺ	1,217	n̄
1,79	đ	1,146	ĺ	1,218	R̀
1,80	Ø	1,147	Ļ	1,219	r̀
1,81	ø	1,148	ļ	1,220	S̄
1,82	Õ	1,149	Ľ	1,221	s̄
1,83	õ	1,150	ľ	1,222	T̄
1,84	Ý	1,151	Ŀ	1,223	t̄
1,85	ý	1,152	Ł	1,224	Y̌
1,86	Ð	1,153	ł	1,225	y̌
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	D’
		1,157	ŉ	1,229	d’
		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	ś	12,76	ṣ
12,1	q̃	12,41	ţ	12,77	ṭ
12,2	ẽ	12,42	ţ	12,78	ṯ
12,3	ĩ	12,43	u̯	12,79	ẓ
12,4	ũ	12,44	B̲	12,84	ẏ
12,5	ũ	12,45	D̲	12,85	á
12,6	ú	12,46	D̲	12,86	í
12,7	q́	12,47	Ĕ	12,87	à
12,8	ẹ́	12,48	G̃	12,88	ì
12,9	ĩ	12,49	H̱	12,89	ù

No.	Character	No.	Character	No.	Character
12,10	$\acute{u}$	12,50	$\underline{H}$	12,90	$\dot{q}$
12,11	$\check{l}$	12,51	$\check{I}$	12,91	$\bar{\delta}$
12,12	$\tilde{m}$	12,52	$\check{J}$	12,92	$\acute{x}$
12,13	$\tilde{r}$	12,53	$\underline{K}$	12,93	x^{v}
12,14	$\hat{r}$	12,54	$\check{O}$	12,94	η^{u}
12,15	$\ddot{r}$	12,55	$\underline{P}$	12,95	$\grave{r}$
12,16	$\hat{a}$	12,56	$\bar{R}$	12,96	$\bar{\partial}$
12,17	$\hat{e}$	12,57	$\underline{S}$	12,97	$\underline{\partial}$
12,18	$\hat{\imath}$	12,58	$\underline{T}$	12,98	$\underline{b}$
12,19	$\hat{o}$	12,59	$\underline{T}$	12,99	$\underline{g}$
12,20	$\hat{u}$	12,60	$\underline{Z}$	12,100	$\underline{h}\upsilon$
12,21	$\ddot{a}$	12,61	∵	12,101	$\acute{k}$
12,22	$\ddot{e}$	12,62	$\underline{b}$	12,102	$\underline{l}$
12,23	$\ddot{\imath}$	12,63	$\underline{d}$	12,103	$\underline{l}$
12,24	$\ddot{o}$	12,64	$\underline{d}$	12,104	$\acute{m}$
12,25	$\ddot{u}$	12,65	$\check{e}$	12,105	$\underline{m}$
12,26	$\acute{e}$	12,66	∂	12,106	$\dot{n}$
12,27	$\tilde{e}$	12,67	$\bar{g}$	12,107	$\acute{\eta}$
12,28	$\tilde{y}$	12,68	$\underline{h}$	12,108	$\underline{n}$
12,29	$\tilde{E}$	12,69	$\underline{h}$	12,109	$\underline{r}$
12,30	$\underline{O}$	12,70	$\check{\imath}$	12,110	$\acute{r}$
12,31	$\underline{o}$	12,71	$\check{\jmath}$	12,111	$\bar{r}$
12,38	$\underline{l}$	12,72	$\underline{k}$	12,112	$\acute{r}$
12,39	$\underline{m}$	12,73	$\check{o}$	12,113	$\check{s}$
		12,74	$\bar{p}$	12,114	$\bar{m}$
		12,75	$\bar{r}$	12,115	$\bar{n}$

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	ח	𐬓	*ϑ*	9,42	ׄ	𐬸	*ū*
9,1	ב	𐬁	*β*	9,20	פ	𐬔	*f*	9,84	בּ	𐬡	*b*
9,2	ג	𐬂	*γ*	9,21	ׂ	𐬕	*j*	9,85	גּ	𐬢	*g*
9,3	ד	𐬃	*δ*	9,22	צ	𐬖	*c*	9,87	דּ	𐬣	*d*
9,4	ה	𐬄	*h*	9,23	ק	𐬗	*ŋ*	9,88	הּ	𐬤	*ą*
9,5	ו	𐬅	*v*	9,24	ר	𐬘	*r*	9,89	וּ	𐬥	*ą́*
9,6	ז	𐬆	*z*	9,25	ש	𐬙	*š*	9,90	זּ	𐬦	*ž*
9,7	ח	𐬇	*x*	9,26	שׁ	𐬚	*š̌*	9,91	חּ	𐬧	x^v
9,8	ט	𐬈	*ţ*	9,31	־	𐬛	*i*	9,92	טּ	𐬨	δ_2
9,9	י	𐬉	*ẏ*	9,32	֑	𐬜	*ē*	9,93	יּ	𐬩	*y*
9,10	ך	𐬊	*ġ*	9,33	֒	𐬝	*e*	9,94	כּ	𐬪	*k*
9,11	כ	𐬋	*x́*	9,34	֓	𐬞	*u*	9,103	מּ	𐬫	*ŋ́*
9,12	ל	𐬌	(*l*)	9,35	֔	𐬟	*ā*	9,104	נּ	𐬬	*ŋ*
9,13	ם	𐬍	*mˌ*	9,36	֕	𐬠	*ā̊*	9,106	סּ	𐬭	*s*
9,14	מ	𐬎	*m*	9,37	֖	𐬡	*ə*	9,107	פּ	𐬮	*p*
9,15	ן	𐬏	*ń*	9,38	֗	𐬢	*ō*	9,108	צּ	𐬯	*ŋ*
9,16	נ	𐬐	*n*	9,39	֘	𐬣	*ī*	9,109	קּ	𐬰	$ŋ^u$
9,17	ס	𐬑	*s*	9,40	֙	𐬤	*å*	9,111	שׁ	𐬱	*ś*
9,18	ע	𐬒	*ā*	9,41	֚	𐬥	*o*	9,114	תּ	𐬲	*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.
𐬓	Ü	25	$\mathring{a}$	182	12,91	9,36	$\mathring{a}$	0101 + 030A	10B03	𐬓
𐬔	ä	22	$ą$	191	1,95	9,88	$ą$	0105	10B04	𐬔
𐬕	Ä	21	$\dot{ą}$	181	12,90	9,89	$\dot{ą}$	0105 + 0307	10B05	𐬕
𐬖							$\bar{ą}$	0105 + 0304		𐬖 19
𐬗	ö	24	$ə$	198	12,66	9,37	$ə$	01DD	10B06	𐬗
𐬘	Ö	23	$\bar{ə}$	199	12,96	9,18	$\bar{ə}$	01DD + 0304	10B07	𐬘
𐬙	e	101	e	101	0,101	9,33	e	0065	10B08	𐬙
𐬚	E	69	$\bar{e}$	167	1,111	9,32	$\bar{e}$	0113	10B09	𐬚
𐬛	o	111	o	111	0,111	9,41	o	006F	10B0A	𐬛
𐬜	O	79	$\bar{o}$	169	1,165	9,38	$\bar{o}$	014D	10B0B	𐬜
𐬝	i	105	i	105	0,105	9,31	i	0069	10B0C	𐬝
𐬞	I	73	$\bar{\imath}$	168	1,133	9,39	$\bar{\imath}$	012B	10B0D	𐬞
𐬟	u	117	u	117	0,117	9,34	u	0075	10B0E	𐬟
𐬠	U	85	$\bar{u}$	170	1,193	9,42	$\bar{u}$	016B	10B0F	𐬠
𐬡	k	107	k	107	0,107	9,94	k	006B	10B10	𐬡
𐬢	x	120	x	120	0,120	9,7	x	0078	10B11	𐬢
𐬣	X	88	$\acute{x}$	183	12,92	9,11	$\acute{x}$	0078 + 0301	10B12	𐬣
𐬤	w	119	x^{v}	184	12,93	9,91	x^{v}	0078 + 036E	10B13	𐬤
𐬥	g	103	g	103	0,103	9,85	g	0067	10B14	𐬥
𐬦	K	75	$\dot{g}$	221	1,125	9,10	$\dot{g}$	0121	10B15	𐬦
𐬧	G	71	γ	223	8,7	9,2	γ	03B3	10B16	𐬧
𐬨	c	99	c	99	0,99	9,22	c	0063	10B17	𐬨
𐬩	j	106	j	106	0,106	9,21	j	006A	10B18	𐬩
𐬪	t	116	t	116	0,116	9,114	t	0074	10B19	𐬪
𐬫	F	70	ϑ	253	8,17	9,19	θ	03B8[37]	10B1A	𐬫
𐬬	d	100	d	100	0,100	9,87	d	0064	10B1B	𐬬
𐬭	D	68	δ	219	8,9	9,3	δ	03B4	10B1C	𐬭
𐬮							$\acute{\delta}$	03B4 + 0301		𐬮
𐬯	T	84	$ţ$	251	12,78	9,8	$ţ$	0074 + 0330	10B1D	𐬯

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

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.

Bibliography

Anquetil-Duperron, M. (1771). Zend-Avesta, ouvrage de Zoroastre, contenant les idées théologiques, physiques & morales de ce législateur, les cérémonies du culte religieux qu'il a établi, & plusieurs traits importants relatifs à l'ancienne histoire des Perses / trad. en François sur l'original Zend, avec des remarques; & accompagné de plusieurs traités propres à éclaicir les matieres qui en font l'objet par M. Anquetil du Perron. Vol. 2, Paris: Tilliard.

Bartholomae, C. (1895–1901). Awestasprache und Altpersisch. In: Grundriss der iranischen Philologie, 1. Bd., 1. Abtheilung, ed. by W. Geiger and E. Kuhn. Strassburg: Trübner. 152–248.

——(1904): Altiranisches Wörterbuch. Strassburg: Trübner.

Chardin, J. (1686). Iournal du Voyage du Chev.r Chardin en Perse & aux Indes Orientales, Par la Mer Noire & par la Colchide, Londres; Journal du Voiage du Chevalier Chardin en Perse & aux Indes Orientales, par la Mer Noire & par la Colchide, Amsterdam: Wolfgangh. English edition: Travels of Sir Iohn Chardin into Persia and ye East Indies Through the Black-Sea And the Country of Colchis, Vol.1. London: Printed for Moses Pitt. [second edition 1691]

——(1711). Voyages de Mr. Le Chevalier Chardin, En Perse, Et Autres Lieux De L'Orient. Enrichi d'un grand nombre de belles Figures en Taille-douce, représentant les Antiquitez & les Choses remarquables du Pais. Vol. 9, Amsterdam.

de Chinon, G. (1671). Relations nouvelles du Levant ou traités de la religion du Gouvernement, & des Coûtumes des Perses, des Armeniens & des Gaures. Lyon: Thioly.

Geldner, K. F. (1889-96). Avesta. The Sacred Books of the Parsis, Stuttgart: Kohlhammer.

Gippert, J. (1995). TITUS. Das Projekt eines indogermanistischen Thesaurus. LDV-Forum 12/2 (1995), 35-47 (also in http://titus.uni-frankfurt.de/personal/jg/pdf/jg1995c.pdf)

Hoffmann, K. (1971). Zum Zeicheninventar der Avesta-Schrift. In: Festgabe deutscher Iranisten zur 2500-Jahrfeier Irans, ed. by W. Eilers. Stuttgart: Hochwacht Druck. 64-73 [repr. 1975 in: Aufsätze zur Indoiranistik. Vol. 1. Ed. by J. Narten. Wiesbaden: Reichert. 316-326, with Table on p. 73 / 326].

Hoffmann, K. and B. Forssman (1996). Avestische Laut- und Flexionslehre. Innsbruck: Institut für Sprachwissenschaft.

Hyde, T. (1700). Historia Religionis Veterum Persarum, Eorumque Magorum. Ubi etiam nova Abrahami, & Mithræ, & Vestæ, & Manetis, &c. Historia, atque Angelorum Officia & Præfecturæ ex veterum Persarum sententia. ... Dantur Veterum Persarum Scripturae & Linguae ... Specimina. Oxonii: e Theatrum Sheldonianum. Cap. XXXV: 'De Persiae & Persarum nominibus; et de modernâ atque veteri linguâ Persicâ ejusque Dialectis; sc. de Persiâ, Parthiâ, Mediâ, & de harum regionum Linguis & Dialectis. 413-428.

——(1760). Veterum Persarum et Parthorum et Medorum Religionis Historia, editio secunda. Ed. by G. Costard. Oxonii: e typographeo Clarendoniano.

Lord, H. (1630). The Religion of the Persees. As it was Compiled from a Booke of theirs, contayning the Forme of their Worshippe, written in the Persian Character, and by them called Zundavastaw. London: T. and R. Cotes.

Pietraszewski, I. (1858). Zend-Avesta ou plutôt Zen-Daschta, expliqué d'après un principe tout à fait nouveau. Vol. 1. Berlin: chez l'auteur éditeur.

Westergaard, N. L. (1852-54). Zendavesta or The Religious Books of the Zoroastrians. Copenhagen: Gyldendal.

Thomas Jügel

Peculiarities of Avestan Manuscripts for Computational Linguistics[1]

Abstract

This paper will discuss several computational tools for creating a stemma of Avestan manuscripts, such as: a letter similarity matrix, a morphological expander, and co-occurrence networks. After a short introduction to Avestan and Avestan manuscripts and a representation of Avestan peculiarities concerning the creation of stemmata, the operatability of the above-mentioned tools for this text corpus will be discussed. Finally, I will give a brief outlook on the complexity of a database structure for Avestan texts.

Introduction

The Avesta, represented by the edition of GELDNER (1886-96), appears to be a sort of Bible containing several books or chapters, cf. SKJÆRVØ's "sacred book of the Zoroastrians" (2009: 44); and, indeed, in Middle Iranian times (i.e., before 600 AD) there existed a kind of text corpus, rather than 'a book', of holy texts (CANTERA 2004). However, GELDNER's edition disguises the actual texts of the manuscripts because what we have today is not a book but a collection of ceremonies attested in various manuscripts.

Avestan is the term for an Old Iranian language, as such a member of the Indo-European language family. The actual name of the language is not known to us. The name 'Avestan' is taken from Middle Persian texts which refer to their religious text corpus as the "abestā(g)". When manuscripts containing these religious texts came to light for European research, they were referred to as "Avesta" and the language as "Avestan".[2]

Avestan is known to us in two varieties, called "Old Avestan" and "Young Avestan". This is so because they display two different chronological layers of Avestan. However, they also differ in some linguistic respect so that they represent two different dialects of the same language (e.g., genitive singular of *xratu-* "wisdom" is *xratāuš* in Old Avestan but *xraϑβō* in Young Avestan, for further examples see DE VAAN 2003: 8ff.).

The Avestan manuscripts (henceforth MS) can be sorted into several groups, the main grouping is: 1) the 'Pahlavi-MSs', and 2) the 'Sade-MSs'. The Pahlavi-MSs contain the Avestan text plus its translation and commentaries, generally Middle Persian, but there are translations into Sanskrit, Gujarati and/or New Persian as well.[3] The Sade-MSs (i.e., the "pure" MS) only contain ritual instructions in Middle Persian, etc., besides the Avestan text. The Pahlavi-MS served as exegetical texts written for scholarly use only. On the contrary, the Sade-MSs were for the daily use in the ceremonies. These different purposes had an influence on the copying process (cf. Section 1).

The aforementioned grouping can be made by first glance at the MS because of the various writings these MSs do or do not contain. Besides the grouping into Pahlavi- and Sade-MSs, the MSs are further classified into different ceremonies. There are four of them: the Yasna Rapihwin, Vīsprad, Yašt, and Vīdēvdād ceremony. Depending on the season or on the deity who is invoked, there are further differences in what is otherwise the same

ceremony.[4] These latter groupings are veiled in GELDNER's edition and this may lead to wrong approaches when it comes to generating the stemmata[5] of Avestan MSs (cf. Section 2).

1 The copying process

If the scribe copies a MS used for the scholarly work on Avestan with all its commentaries (i.e., a Pahlavi-MS), the main interest is to produce the exact copy of the original. Together with the MS itself, the colophon is also copied, since it serves as a kind of proof of quality when the list of authorities is given. In this process the original was usually not corrected, probably not even read (if the scribe could read Avestan at all). So loss of lines, even of pages, often went unnoticed (cf. CANTERA 2010). Furthermore, it might well be the case that the scribe has mixed different styles of writing. There are, e.g., two versions of the characters ⟨ą⟩, i.e., nasal /a/, and of ⟨y⟩.[6] The one is typical for Iranian MSs, the other one for Indian MSs. We know that Iranian MSs were brought to India. An Indian scribe copying an Iranian MS would have had the choice of copying not just the text but also the style of writing or of converting the Iranian features into Indian ones. This transfer would surely not take place consistently; and, indeed, some MSs show both features.

The scribe who is copying a Sade-MS, which is used in everyday life, would want to produce the best text, not the best copy. As there might have been scribes who could not read Avestan and Middle Persian very well, others were surely experts in it, having a high knowledge of the Avestan ceremonies as well. In the tradition of these MSs, loss of text was usually noticed and the text restored – not always with the correct result as we can say today. A telling disimprovement on the word level was presented by CANTERA on occasion of the conference "Poets, priests, scribes and librarians: the transmission of the holy wisdom of Zoroastrianism" (Salamanca 2009):

> Frequently a final *-əṇg* appears in the manuscripts as *-əṇga*, clearly a reflection of the pronunciation with a final epenthetic vowel. Since this error was known to the priests, they sometimes made hypercorrections. Thus the well-known Indian scribe Dārāb Hīrā "corrected" the right *vīspāṇg. āiiōi* in Y31.2 into the wrong *vīspāṇg. ẏōi*. He obviously thought that *ā* was an epenthetic vowel in the pronunciation of final *-əṇg*. Such hypercorrections occur often.

Here, the scribe took the prefix of the verb *āiiōi* "I ask for" as the epenthetic final vowel of the preceding word. That it was written separately does not raise objections to this wrong analysis since this would be normal for enclitics. However, the reanalysis goes further. The remaining °*iiōi* was understood as the relative pronoun *yōi* (nominative plural masculinum), and ⟨ii⟩ is the orthography of non-initial /y/. Thence °*iiōi* was changed into *yōi*. Another of such erroneous reanalyses is *aēšəm. mahiiā* in Yasna 48.12. The genitive singular of *aēšəma-* "fury", i.e. *aēšəmahiiā*, was split up into two words: *aēša-* "capable" and the genitive singular of the possessive pronoun *ma-* "my". This reanalysis seems to be very old because all of the MSs known to us show some variation at this point.[7]

Besides such corrections, there also occurred alternations of the ceremony, whether because the scribe was following a different custom, or because he had heard of a variant

that he considered better because it was being propagated by a high authority (*mobile variants*).

Obviously, the tradition of Pahlavi-MSs and Sade-MSs proceeded quite differently. We do not expect the same peculiarities of copying processes for both of these groups. The *mobile variants* also blur the border of the various types of ceremonies. Hence, a noteworthy alternation might not be due to the manuscript's belonging to the same group, but rather to external influence on the copying process.

2 The difficulties in generating a stemma for Avestan manuscripts

A huge part of the Avestan corpus is lost. We know this because there are references in the Middle Persian Zoroastrian literature to Avestan passages which were not passed down. Furthermore, the majority of copies is not in our reach – either because they were lost, or because their whereabouts are unknown to the scientific world. There must have been a time when plenty of copies were produced in a year. Some colophons were written by one hand including the year, but the name of the copyist was added later by another hand (CANTERA 2012: 298). There were families whose profession seemed to have been the production of Avestan manuscripts. The ADA project has located more than 300 MSs so far. So, we can consider ourselves extremely lucky whenever we find the rare case of having both the mother MS and its daughter MS or the direct siblings of one mother MS. Such cases show, by the way, that the differences of copies by one and the same author can be much higher than differences of copies from a different copyist (as is the case with the MSs K1 and L4, cf. CANTERA 2012: 329). That is, some copyists did not work very accurately.

Apart from the scarcity of the remaining MSs, we have to consider the impact of the copying process, as described in Section 1, such as deliberate emendations or *mobile variants*. Hence, concentrating on variants which manuscripts may have in common can lead to a distorted picture as it is the case in GELDNER's prolegomena.

The relationship of manuscripts is not necessarily the same as the relationship of the text/textual variants. Manuscripts can be dated according to colophons or by analysing the material (paper, ink). A single MS can be split up into chronological layers when there are emendations and additions of a second hand, which may reveal the influence of another *vorlage*. The text, however, is more abstract. It is *a priori* not clear whether the text of a MS of the 18th century is indeed younger than the text contained in a MS of the 16th century, as aptly put by MINK (2004: 24, italics original): "*the text is the witness, not the manuscript*".

The "Coherence-Based Genealogical Method" (CBGM)[8] combines computational means with philological know-how. At first, the comparison of MSs leads to a so-called "pre-genealogical coherence". A high similarity between manuscripts speaks in favour of a close relationship. Then, for each significant variation a local stemma is philologically stipulated resulting in a textual flow, i.e., each MS is put in chronological relation with the others. An arrow between two MSs in the textual flow does not mean that MS *b* was copied from MS *a*, rather that text *b* is in the textual flow a younger witness being somehow influenced by MS *a*. Furthermore, in a given textual flow a MS can have more than one arrow pointing to it, since its text may have been influenced by the one of several MSs (e.g., in cases of *mobile variants* or collocations from various manuscripts). In order to represent the

various degrees of influence the arrows show different widths, i.e., they are substituted by vectors (CANTERA 2012: 320). Combining the degree of pre-genealogical coherence with the textual flow and the local stemmata yields a global stemma (or stemmata if several sub-stemmata are equally possible), cf. the example in Figure 1 and its discussion thereafter.

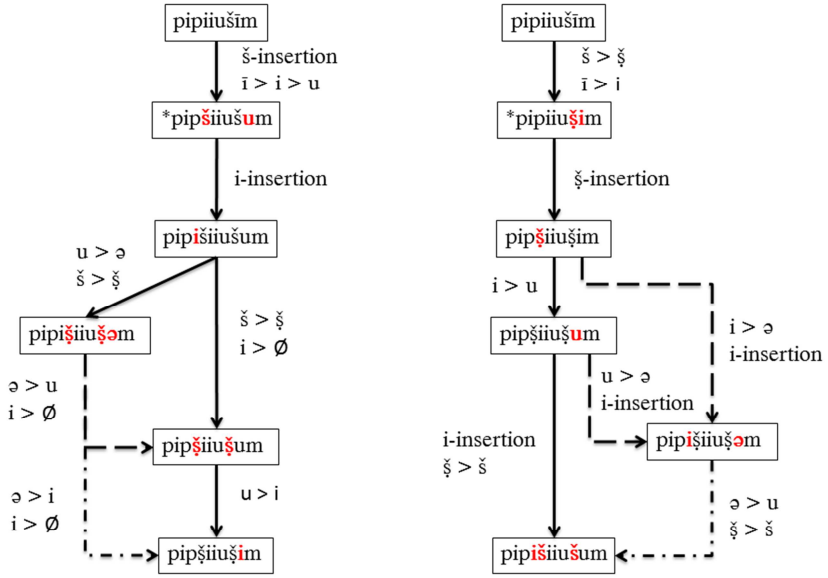


Figure 1: Two possibilities of building a local stemma (data taken from CANTERA 2012: 341)

The original form is *pipiiūšīm* "swollen", a feminine *i*-stem in the accusative singular. In the MSs appear several variants: A) *pipišiiušum*, B) *pipişiiuşəm*, C) *pipşiiuşum*, D) *pipşiiuşim*. There are two equally likely possibilities of how the variants could be arranged in a chain of derivation.

The various ⟨š⟩ characters were confused in the MSs (cf. Section 3.1) so that we may stipulate an intermediary form **pipiiūşīm* (right branch): another ⟨ş⟩ was added by mistake. The length of /i/ and /u/ are not always kept distinct so that the D-variant *pipşiiuşim* may readily evolve. The letters representing ⟨i⟩ and ⟨u⟩ can easily be confused due to their similarity in form, so that we get the C-variant *pipşiiuşum*. In order to explain the B-variant *pipişiiuşəm*, we apply an orthographic rule, viz., if a consonant is followed by ⟨i⟩ or ⟨ii⟩, another ⟨i⟩ is written in front of it (presumably indicating the consonant's palatal pronunciation). The /ə/ might be a phonological reduction of either /i/ or /u/, respectively, revealing an influence of the scribe's pronunciation. For the A-variant *pipišiiušum*, the insertion of an

orthographically motivated ⟨i⟩ is stipulated and, again, confusion of the various characters of ⟨š⟩. A derivation of the A-variant from the B-variant is less likely because /ə/ does not easily change into /u/ (though one could imagine assimilation to the labial in the penultimate syllable). The left branch is equally possible with the same explanations just differently ordered. While in the right branch ⟨š⟩ changes via ⟨š̨⟩ back to ⟨š⟩, in the left branch ⟨ī⟩ changes via ⟨i⟩ and ⟨u⟩ back to ⟨i⟩.

If we take the textual flow into account, we see that the MSs showing the A-variant are generally prior to those containing the B-variant, which are prior to the C- and D-variants, i.e., the left branch of the local stemma is probably the correct one.[9]

Applying CBGM to Avestan is extremely labour-intensive. One has to digitize manuscripts and to detect and evaluate variants. A high degree of philological knowledge is vital for the evaluation. In order to accomplish such an ambitious task, several scientists of European institutions have agreed on a cooperation which celebrated its constitution as "Corpus Avesticum" on occasion of a workshop held in Frankfurt am Main in November 2011.[10] The work of the philologist can be facilitated by means of computational devices. The following sections discuss their pros and cons.

3 The Avestan Language and Computational Devices

3.1 Simulation of the copying process – interchangeability of characters

In order to set up a local stemma, words at a variant position are aligned in such a way that characters in corresponding positions in sample form a pair of characters. For instance, assuming that the words *aiiese* and *aiesē* would occur as possible variants, the first pair of characters would be "*a-a*". If the variants differ in the number of letters, then a gap is inserted into the shorter word and aligned with the corresponding letter of the longer one (e.g., *a-a*, *i-i*, *i-Ø*, *e-e*, *s-s*, *e-ē*). A distance function sums up the distance values of each pair of characters of a variant word pair, normalizes them by dividing by the mean length of the two words, and returning this as an overall value of their distances. With an alignment done by established measures such as the Levenshtein distance (LEVENSHTEIN 1966), each difference in characters would have the same weight.

However, scribes did not randomly substitute one character by another (e.g., writing ⟨k⟩ instead of ⟨a⟩), but rather they were following certain logical rules. Either characters could easily be misread (e.g., ⟨ī⟩ and ⟨ū⟩ in Avestan script), or, according to their phonological surroundings, sounds could be confused and, hence, the characters which represent these sounds (e.g., /a/ and /e/ in a palatal context despite the shapes of these two characters being otherwise clearly distinct). In order to be able to tell trivial changes from non-trivial ones, an Avestan-specific two-dimensional matrix of the interchangeability of characters had to be stipulated (Figure 2 below).[11] The differences of the pair of characters are weighted by applying this lookup table, or matrix of distances, by a distance function. In Figure 2, one dimension characterizes the likelihood of two characters being exchanged due to phonological (and rarely also orthographic) reasons as likely (green), possible (white), unlikely (red). For example, in Old Avestan, final vowels were always long, while in Young Avestan they were always short (with the exception of monosyllabic words and a few flectional endings).

This was, of course, apparent to the scribes as well, who may have tried to archaize Young Avestan texts. So the difference of final *-a* and final *-ā* may simply be of no importance. In palatal context vowels may have been palatalized so that in a sequence like *-iiami-* the variant *-iiemi-* is of little significance.[12] The difference of the sound represented by the characters ⟨β⟩ and ⟨uu⟩ (i.e., bilabial /w/) is a phonetic one, not a phonemic one. The word *ϑβaršta* may also appear as *ϑuuaršta*.

a	**ā**	**å**	**å̄**	**ą**	**ą̇**	**ə**	**ə̄**	**e**	**ē**	**o**	**ō**	**i**	**ī**	**u**	**ū**	
x	9	8	7	1	1	1	1	1	1	1	1	3	1	1	2	**a**
	x	4	8	1	1	1	1	1	1	1	1	3	1	1	1	**ā**
		x	9	1	1	6	5	1	1	1	1	3	1	1	1	**å**
			x	1	1	6	5	1	1	1	1	3	1	1	1	**å̄**
				x	9	3	2	1	1	1	1	1	1	5	1	**ą**
					x	1	1	1	1	1	1	1	1	1	1	**ą̇**
						x	9	1	1	1	1	1	1	4	1	**ə**
							x	1	1	1	1	1	1	2	1	**ə̄**
								x	8	1	1	1	1	1	1	**e**
									x	1	1	1	1	1	1	**ē**
										x	9	1	1	1	1	**o**
											x	1	1	1	1	**ō**
												x	9	8	3	**i**
													x	2	9	**ī**
														x	8	**u**
															x	**ū**

Figure 2: matrix of the interchangeability of vowels

The other dimension of the matrix represents the grade of similarity of the shapes of characters. The higher the figure (1-9), the more similar the characters. The characters ⟨δ⟩ and ⟨γ⟩, ⟨y⟩ and ⟨š́⟩, or ⟨ī⟩ and ⟨ū⟩ can easily be confounded, though linguistically it is rather unlikely. Aside from the comparison of single characters, character groups also have to be compared. There is a high similarity of ⟨ai⟩ and ⟨ā⟩, ⟨šk⟩ and ⟨š̌⟩, or of ⟨an⟩ and ⟨x^{v}⟩, etc. Phonetically, the difference of ⟨ŋuh⟩ and ⟨ŋ^{v}h⟩ is lacking since both are just two different ways of expressing the same sound: a labialized laryngeal with a nasalizing effect on the preceding vowel (cf. HOFFMANN/FORSSMAN 2004: 45).

There is one striking instance of interchangeability that is not due to the high similarity of the shapes of the characters or of the sounds the characters represent but rather to orthographic conventions.[13] This concerns the characters ⟨h⟩, ⟨s⟩, and ⟨ϑ⟩. All three sounds these characters represent existed in the Old Iranian languages Avestan and Old Persian. In Middle Persian, however, /ϑ/ changed to /h/. In the cryptic orthography of Middle Persian, an /h/ could be indicated by the characters ⟨h⟩, ⟨s⟩, and ⟨t⟩: ⟨h⟩ for wherever it is the normal

representation of /h/, ⟨s⟩ in non-Persian words wherever non-Persian /s/ equalled Persian /h/ due to the results of sound change, and ⟨t⟩ wherever it was the archaic representation of /θ/, which later became /h/. Given all this, when native speakers of Middle Persian pronounced Avestan, they could have substituted /θ/ by /h/ and written it accordingly, or they might simply just confused the three characters ⟨h⟩, ⟨s⟩, and ⟨θ⟩ due to Middle Persian orthographic conventions. In fact, there are only few instances of ⟨s⟩/⟨θ⟩ confusion.[14]

A sporadic variant that is due to the confusion of characters of high similarity or of similar sounds is not significant for the grouping of MSs. Such confusion could have happened at any time. However, if there is a high regularity of such unspecified changes, they might be telling nevertheless.

A programme able to apply the matrix described above can produce a distribution of weighted letter substitutions and will help the philologist to concentrate on relevant variants that allow the stipulation of a local stemma. Those variants that are due to trivial changes will only be evaluated when they are needed for the constitution of a local stemma that comprises significant changes.[15]

3.2 Morphological expansion

The automatic generation of paradigms is helpful for text technological analyses of word forms (e.g., POS) that have not been entered in the digitized lexicon. Therefore, it is necessary to feed the programme all information needed, e.g., sets of endings, inflectional classes, stem alternations, etc. In highly standardized languages such a task can be accomplished with reasonable effort. In Avestan, however, it is much more complicated. To begin with, there is no standardized orthography; the orthographical conventions are rather tendencies. So the rules for the interchangeability of characters described in Section 3.1 have to be applied to the analysis of word forms as well. Then, most of the nominal suffixes have to be entered in as two variants because there is a differing output of endings in so-called *sandhi* context: while, e.g., word final **-as* (ending of the nominative singular masculinum) developed via **-ah* to *-ō* in Avestan, **-as* followed by the enclitic *-ča* "and", i.e., in sandhi context, was preserved (so there is *haomō* besides *haomasča*). Strictly speaking, the paradigm of each declension should show each ending in pausa as well as in sandhi context.

The combination of suffixes may also lead to a different phonological output. So, one cannot simply combine them. For example, the suffix **-ant-* has two different outputs: 1) *-əṇt-*, 2) *-ąt-*. Whether the one or the other output is to be expected depends on the phonological surrounding, i.e., which suffix is following.[16]

Sometimes it is hard to tell whether the variants shown by the MSs are linguistic variants due to dialectal or chronological differences, or whether they are the result of the copying process. However, even in the cases where we do know the regularities, they are so numerous that the task of installing grammatical rules for automatic generation seems hardly worth the effort. As an example I shall explain the paradigm of *pitar-* "father" in detail:

- nominative singular: *ptā* besides *tā* and *pita*
- accusative singular: *patarəm* besides *pitarəm*
- dative singular: *fəδrōi* besides *piθrĕ*

Seeing this irregular paradigm, we may say that "luckily" not more forms are attested. The Indo-European nominative singular **ph$_2$térs* developed as follows: The laryngeal **h$_2$* was either lost or vocalized to *i*. The vowel **e* regularly changed to *a*, and the auslaut **rs* was assimilated to **r* plus compensatory lengthening of the vowel (Szemerenyi's law), followed by a yet unexplained loss of the final **r*. This yields the Iranian output *pitā*, or *ptā*, respectively. The uncommon onset *pt* was simplified to *t*, i.e., *ptā* > *tā*. The Young Avestan form shows the shortening of final vowels, hence *pita*. The Indo-European accusative **ph$_2$térm* developed via **p(i)taraəm* to either *pitarəm* or *ptarəm*, where an anaptyctic /a/ was inserted in the later tradition of Avestan. The Indo-European dative **ph$_2$tréi̯* developed via **fθraı̯* to *fθrōi*, which displays the Old Avestan development of **ai̯* to *ōi* at the end of a word. Again, an anaptyctic vowel /ə/ was introduced. Besides these irregular forms, there was an analogically introduced stem *piθra-*, which displays the fricativization of preconsonantal voiceless stops (i.e., **tr* > *θr*). The form *piθrĕ̄* shows the Young Avestan output of word final **ai̯*.

So what synchronically seems like a nightmare for every child or non-native speaker learning this language, can easily be explained by the linguist from a diachronic point of view. The rules, however, do not outweigh the irregularities that are due to phonological effects (sound change), analogical formations (morphological effect), or to reflexes of spoken language (assimilation in the course of recitations).

When it comes to a language like Avestan with such a small corpus of less than 12920 words (DOCTOR 2004: 5),[17] it is easier to annotate every single form by hand. The automatic production of non-attested word forms would always remain highly hypothetical and offer very little insight. Nevertheless, a morphological expander is helpful in suggesting to the philologist the most likely form.

The irregularity not only affects declension or conjugation but also the stems themselves, i.e., not only the grammar but the lexicon as well. The word *napāt-* "grandson" (cognate to English 'nephew') is attested with three different stems: *napāt-*, *naptar-*, and *napa-*. These alternations are not simple mistakes. They are of high interest and show some linguistically well-known patterns. *napāt-* is the inherited form. *naptar-* is a transmutation in analogy to other words denoting family terms like *pitar-* "father", *brātar-* "brother", *mātar-* "mother", etc. This change is due to the semantic class of family terms, most stems of which end in °*tar-*, hence *napāt-* > *naptar-*. The stem *napa-* is based on a regularization process. The many declensional classes of Old Iranian were simplified to a few, the most dominant one being the *a*-declination. Words were extended by *-a-* to make them fit into this class, e.g., *n*-stem *zruuan-* "time" besides the newly formed *a*-stem *zruuāna-*. In the case of *napāt-* the form was shortened to *napa-*. The patterns described are not a sign of degeneration but of language development along the lines of logical reasoning. So, we do not want to emend this. We want to find it. Therefore, each variant gets its own entry in the lexicon representing an inflexion class of the lemma (cf. LINDE this volume).

3.3 Co-occurrences and citation

A handy tool of digitized corpora is a co-occurrence analyzer to check the contexts of a word, i.e., it yields those words occurring with the target most frequently. Therefore, a window is defined comprising the given word and its neighbours. In manuscripts which

contain interpunctuation marking clause boundaries the frame usually is the sentence. Where one cannot detect such boundaries, a sequence of words containing the target is taken instead. In languages rich in morphology, this sequence may be smaller than in those with an analytic system. For instance, German only exhibits four cases (nominative, accusative, genitive, dative) the marking of which are many times syncretistic, i.e., the same suffix is used for different cases. Several nouns only distinguish number (singular, plural) and are not marked for case at all (e.g., *Frau* "woman", *Frauen* "women"). It is the preceding article that makes case forms distinguishable (e.g., *die Frauen* nominative/accusative, *der Frauen* genitive, *den Frauen* dative, all plural). Hence, such languages like German display a huge amount of functional words (articles, auxiliaries, adpositions, and particles). Languages with a rich inflectional system like Avestan do not need these functional words. Instead, they show longer words exhibiting all kinds of functional information in the affixes.[18]

When the window is defined and a query is made, the result will show the word's significant co-occurrences, which can be filtered by their part-of-speech. Such co-occurrences represent valuable information for historical semantics.[19] Tools such as Linguistic Networks[20] allow the visual representation of co-occurrence networks, i.e., not only the target and its co-occurrences are listed but the co-occurrences of the latter ones as well (cf. the following screenshot, Figure 3).

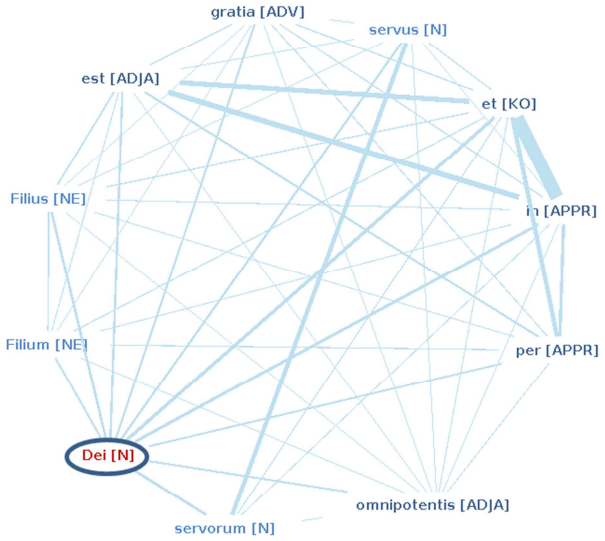


Figure 3: co-occurrences of Latin *dei* "of god"

For Avestan studies, such a query could reveal a differing usage of words in Old and Young Avestan. If Middle Persian is taken into account as well, an alteration of concepts may become visible.[21] For instance, the word *daēnā-* means "religion" in New and Middle Persian (*dēn*, or *dīn*, respectively). However, in Middle Persian another meaning is still detectable. It is the personification of the good or bad deeds of a human. If a man was good, he could expect to meet a beautiful girl in the afterlife who accompanies him to paradise. If he was evil, an ugly, stinking old woman would await him. In the Avestan ceremonies the priest may have contact to the transcendent world and meet his *daēnā-*. The original meaning of *daēnā-* in Avestan is considered to be "view, conception", and not "religion", which has become the traditional translation. Another word of interest is *frauuaši-* "choice", later a personification of the good choices of the ancestors, a guarding spirit.

Avestan texts are the holy texts of Zoroastrianism. However, it is the Middle Persian corpus that is the biggest among the Zoroastrian library. Indeed, we have more texts on Zoroastrianism than holy Zoroastrian texts. Middle Persian texts reveal that there once was a so-called 'Great Avesta' with Middle Persian translation. The Avestan ceremonies we know of today were not necessarily part of this Great Avesta. They may be the textualisation of the spoken ceremonies, i.e., of the practice. Having said this, scientists think that some Middle Persian translations were nevertheless taken from the Great Avesta because they differ in style and translation technique from those that were probably translated directly from the textualised ceremonies. There are few texts that are said to have been part of the Great Avesta (e.g., the Nērangestān, a Middle Persian text with Avestan quotations). If we link the Avestan words, phrases, and clauses with their Middle Persian counterparts, queries will allow classifying and sorting translation techniques, which may differ from text to text. With this knowledge it is then possible to detect Avestan *vorlages* of Middle Persian texts, the Avestan original of which is not known to us. The picture which emerges from such an investigation will show how far Avestan was known to the Zoroastrians of post-Sasanian Persia, i.e., after the Arabic conquest and the spread of Islam. Furthermore, we will get a glimpse into the literary corpus of Sasanian Persia. Even purely Middle Persian texts such as the Bundahišn, an encyclopaedic work, may be based on Avestan *vorlages*. The Avestan Vīdēvdād comprises a legend on the creation of several countries, quite similar to the style of the Middle Persian Bundahišn (chapter 31). So, how Avestan indeed is the Middle Persian corpus?

3.4 Interdependencies of Avestan texts

This section will deal with the complexity of a database the purpose of which is to represent the entire Avestan corpus. The information contained in Avestan manuscripts is allocated to several interdependent segments. As a basis, we can take the Avestan text. Then there are additions and emendations of the text written in the margins or between the lines. These may result directly from the reading or understanding of the Avestan text(s). In the first case, the comparison is drawn to the text committed to memory, which the copyist uses in daily ceremonies. In the second case, these interpolations may result from the Middle Persian translation, which presents yet another layer. Further translations (like into Gujarati) might be based directly on the Avestan text, but are more likely derived from a Middle Persian trans-

lation. So, we can build a hierarchy of dependency. However, an interpolation can bypass an intermediate level and affect a much lower or higher one, e.g., based on the Sanskrit translation of the Middle Persian translation that is the direct translation of the Avestan text, a copyist may decide to "correct" the Avestan text. Besides translations, we also find commentaries that are definitely based on the understanding of the text. These commentaries show influences of the current *zeitgeist* and may have been reinterpreted quite differently by copyists of later centuries. Such reinterpretations – although not changing the wording of the commentary itself – may have had an effect on translations into other languages or, again, may have lead to interpolations of the Avestan text. Avestan text passages are quoted or referred to in Middle Persian texts (e.g., the *Pursišnīhā* "The catalogue of questions"). Although these relations lead outside of the Avestan text corpus itself, viz., to Middle Persian texts, they may reveal the current understanding of the Avestan text at the time the Middle Persian text was written.

The Avestan text itself may be segmented into the Old and Young Avestan texts. There are references to Old Avestan in Young Avestan, and Young Avestan features appear in Old Avestan text segments. The many repetitions – sometimes with small variations and/or short additions – form another set of segments.

Furthermore, there are different ceremonies that only partly display the same text. Variations that belong to different ceremonies could have been judged by copyist as "better" forms (cf. Section 1 "*mobile variants*").

So, we have several layers, some of them arranged horizontally, others vertically, and still others standing in an interdependent relationship. As a corpus is built up step by step, i.e., layer by layer, the interdependency grows and should be taken into account by tools organizing and evaluating the data.

4 Conclusions

I hope to have shown the peculiarities of a language that is only known by its textual sources. Generally, these observations hold to be true for all languages that have not yet developed a standard written form. Avestan is all the more complicated because its oral tradition (later by non-native speakers), its late textualisation, and its subsequent textual tradition led to many effects for which we first need to determine linguistic relevance. Often, we simply still do not know the correct reading of the original. To overcome this, a stemma has to be established. The international scientific cooperation Corpus Avesticum will apply the CBGM method, which combines computational methodology with philological expertise. Several tools will facilitate the work of the philologist, e.g., a tool for finding significant variants by means of a distance function, including a matrix of character interchangeability.

The confusion regarding the original text has prevented comprehensive syntactic studies so far – a job that can easily be accomplished via a well-organized database. Queries on semantics by means of a co-occurrence analyzer will help to elucidate the meaning of unknown words and the development of concepts. Quotation analyses will help to trace the literary-historical development of Avestan and, especially, Middle Persian. These simple tasks require complex spadework: the linking of Avestan with the secondary languages into which it was translated.

An anticipated hurdle is the development of an interactive database that will be available online. Various subcorpora, e.g., a database of manuscripts including their images and metadata, a concordance of digitized texts of the manuscripts, a collection of edited translations, and a database of quotations all need to be interlinked by means of modules such as attestation, lexicon, and grammar. The user should be able to navigate easily from one corpus to the other, or to call up a visualization illustrating how these are linked. Furthermore, the user should have access to adjust, add, and alter information with real-time effect on the linked modules.

Besides the high linguistic and cultural impact of Avestan and the importance of its understanding, the specific problems that the small Avestan corpus presents may motivate us to develop methods and tools that would be useful for other tasks as well.

---

[1] I would like to thank Prof. ALBERTO CANTERA of the University of Salamanca, who so willingly shared his knowledge of Avestan and Avestan stemmatology.

[2] A more detailed survey is to be found in CANTERA (2004).

[3] Note that Indian scripts are dextrograde (left-to-right), while Iranian scripts (ultimately derived from the Aramaic script) are sinistrograde (right-to-left). When it comes to the use of both on the same piece of paper, the scribe faces the problem of organizing the lines in order to avoid one script overwriting the other. This has been prevented by either leaving the rest of the line blank, by jumping into the next line as soon as the dextrograde script reaches the end of the sinistrograde passage (e.g., MS G10), or by turning the leaf 180°, in order to write the dextrograde script upside-down so that it becomes sinistrograde when turned back (e.g., MS S1).

[4] For a more detailed survey on the various types of Avestan MS see CANTERA (2012).

[5] A stemma is a family tree of manuscripts which shows the relationships of the surviving witnesses of a text. Traditionally, each stemma has at its top one original text version.

[6] For a survey on Avestan characters and their encoding see GIPPERT (this volume).

[7] Cf. GELDNER (1886-96: 171 of the Yasna). I re-checked all MSs available on ADA (http://ada.usal.es/paginas/buscador_obra, 3rd April, 2012). Only G18b is with *aēšəm.ahiiā* very close to the original *aēšəmahiiā*. According to TITUS (http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/yasna/yasnavar/yasna.htm), the same is true for the MSs Br2, Jm2, and Jm3.

[8] The CBGM was developed by MINK (cf. 2004 with further references) and adapted to Avestan by CANTERA (2012).

[9] See CANTERA (2012: 341) for a more detailed discussion of this problem.

[10] http://corpusavesticum.hucompute.org/. Constituting members are affiliated to the Universities of Berlin, Bologna, Frankfurt/Main, Göttingen, London, and Salamanca.

[11] For a discussion on letter identification see MUELLER/WEIDEMANN (2012) with further references. – Since the object of our study lies in the past, experiments for stipulating the matrix were impossible. Instead, I knowingly set up the matrix based on my experience and intuition.

[12] It is not always clear when such a difference is due to the pronunciation (probably even by non-native speakers of Avestan) in the recitation and when such changes are the result of sound changes, i.e., a feature to tell dialects or chronological layers apart. Cf. DE VAAN's

(2003: 266f.) discussion of **-čam*, **-ǰam*, **-čam* > **-čim*, **-ǰim*, **-yim*. DE VAAN postulates an intermediary **-čəm*, etc., which is partially preserved in Old Avestan. The development of *ə* > *i* is considered by him to be an effect of "the post-archetype pronunciation", i.e., not a linguistic feature of the language Avestan itself.

[13] Other orthographic conventions, like Indian ⟨y⟩ for Iranian ⟨ẏ⟩, also fulfil the condition of high phonological similarity (in this case both characters represent the same sound). Hence, *ẏąm* vs. *yąm* do not represent two different words or word forms. They are considered to be not two variants but two readings of one variant (CANTERA 2012: 329).

[14] One such example is *ϑrāzdūm* in Yasna 34 §7, which is represented by *srāzdūm* in the MSs Mf1, K37, and Pd (GELDNER 1886-96: 125 of the Yasna), in Br2 (http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/yasna/yasnavar/yasna.htm), and in ML15284 (http://ada.usal.es/paginas/ver/15806).

[15] Confer HOENEN (forthc.) for a theoretical survey on such a programme.

[16] Confer DE VAAN (2003: 624ff.) for rules. For instance, ⟨a⟩ has 12 different inputs, ⟨ā⟩ 13, ⟨ą⟩ 6, ⟨ə⟩ 7, etc.

[17] The number of 12920 words comprises word forms as well, i.e., not only lemmata. However, since DOCTOR also gives compound components as extra entries, the number should be reduced because the first unit of a compound usually does not represent a part-of-speech in its own right. The form is often the stem, or a specific interfix emerges.

[18] The same holds true for agglutinative languages like, e.g., Turkish. In the phrase *bunu yapabileceğinizi söylediniz* "you said that you will be able to do this", the single word *yapabileceğinizi* consists of the following entities: *yap-* stem "to do" + *-abil-* "to be able" + *-eceğ-* for future reference + *-iniz-* "you" (plural) + *-i* for the accusative. That is, what English renders with seven words (*that you will be able to do*) is expressed by a single one in Turkish. This simple example shows that when it comes to comparing languages with one another, language specific features must be taken into account so that whatever is compared (e.g., word length) is indeed comparable.

[19] A respective study is *Virtus. Zur Semantik eines politischen Konzepts von Augustinus bis Johannes von Salisbury*, by Silke Schwandt (PhD-thesis, Frankfurt am Main 2010).

[20] http://www.hucompute.org/ressourcen/linguistic-networks.

[21] Such a study undertaken with classical philological methods is KÖNIG (2010).

Bibliography

Cantera, A. (2004). Studien zur Pahlavi-Übersetzung des Avesta. Iranica 7, Wiesbaden: Harrassowitz.

——(2010). Lost in transmission: The case of the Pahlavi-Vīdēvdād manuscripts1. In: Bulletin of the School of Oriental and African Studies 73/2, 179-205.

——(2012). Building Trees: Genealogical Relations Between the Manuscripts of Videvdad. In: The Transmission of the Avesta, Iranica 20, Wiesbaden: Harrassowitz. 279-346.

Doctor, R. (2004). The Avestā: A Lexico-Statistical Analysis (Direct and Reverse Indexes, Hapax Legomena and Frequency Counts). Acta Iranica 41.

Geldner, K. F. (1886-1896). Avesta – The Sacred Books of the Parsis. Stuttgart: Kohlhammer [reprint of 1381/2003. Tehran: Asāṭīr].

Gippert, J. (2000). Indoiranisches Text-Retrieval. Elektronische Bearbeitung altiranischer und vedischer Texte. In: Indoarisch, Iranisch und die Indogermanistik – Arbeitstagung der Indogermanischen Gesellschaft vom 2. bis 5. Oktober 1997 in Erlangen, ed. by B. Forssman and R. Plath. Wiesbaden: Reichert. 133-145.

——(2002). The Avestan language and its problems. In: Proceedings of the British Academy 116, 165-187.

Hoenen, A. (forthc.). Letter Similarity. In: Proceedings of the Workshop 'Methods and Means for Digital Analysis of Ancient and Medieval Texts and Manuscripts'. Leuven.

Hoffmann, K. and B. FORSSMAN (2004). Avestische Laut- und Flexionslehre. Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

Jacobs, A. M. et al. (1989). Perception of lowercase letters in peripheral vision: A discrimination matrix based on saccade latencies. In: Perception & Psychophysics 46/1, 95-102.

König, G. (2010). Geschlechtsmoral und Gleichgeschlechtlichkeit im Zoroastrismus. Iranica 19. Wiesbaden: Harrassowitz.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics-Doklady 10. 707-710.

Mink, G. (2004). Problems of a highly contaminated tradition: the New Testament – Stemmata of variants as a source of genealogy for witnesses. In: Studies in Stemmatology II, ed. by P. van Reenen et al. Philadelphia: John Benjamins.

Mueller, S. T. and C. T. Weidemann (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. In: Acta Psychologica 139/1, 19-37.

Skjærvø, P. O. (2009). Old Iranian. In: The Iranian Languages, ed. by G. Windfuhr. London/New York: Routledge. 43-195.

DE VAAN, M. (2003): The Avestan Vowels. Amsterdam; New York: Rodopi.

http://ada.usal.es/paginas/buscador_obra

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

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

Roland Mittmann

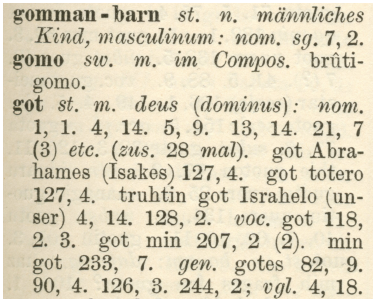
Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen

Um Wörter und Wortformen innerhalb von Texten auffindbar zu machen, waren im vordigitalen Zeitalter Glossare unerlässlich. Heute lassen sich ihre Daten automatisiert mit den zugehörigen Texten zusammenführen, um die Texte so mit weiteren Informationen anzureichern. Für die dazu notwendige Digitalisierung der Glossare ist angesichts des historischen Druckbildes und der oft nicht eindeutigen Informationsauszeichnung ein manuelles Vorgehen am zielführendsten. Je nach Strukturierung des Glossars und nach Art und Überlieferungsdichte des behandelten Textes ergeben sich dabei unterschiedliche Herausforderungen und Probleme. Diese werden am Beispiel der Digitalisierung der Glossare zum Althochdeutschen und Altsächsischen dargestellt.

1 Die Problemstellung

Das digitale Zeitalter hat die technischen Voraussetzungen dafür, automatisch durchsuchbare Textkorpora zu erstellen, mit sich gebracht. Während die Digitalisierung der bloßen Texte heutzutage meist kein Problem darstellt, bleibt weiterhin die Frage, wie sich diese unannotierten Korpora unter möglichst geringem Aufwand mit zusätzlichen Informationen versehen lassen.

Zu vielen altüberlieferten Texten sind im 19. und 20. Jahrhundert Glossare erstellt worden, mit Hilfe derer es möglich ist, zu einem bestimmten Lemma dessen belegte Wortformen zu eruieren und die Stellen im Text zu ermitteln, an denen diese Wortformen erscheinen (vgl. Abbildung 1). Diese Glossare könnten somit also eine ergiebige Wissensbasis darstellen, wenn es durchführbar ist, im Zuge der Automatisierung eine Suche in anderer Richtung vorzunehmen: Hierzu müsste man von jedem einzelnen im Text belegten Wort ausgehen, diesen Beleg im Glossar wiederfinden und bei Mehrdeutigkeit zur eindeutigen Zuordnung auch die Position im Text abgleichen, dann die Angaben zu morphologischen Eigenschaften der Wortform auslesen und schließlich auch das Lemma sowie wiederum die zum Lemma angegebenen morphologischen Informationen extrahieren. All diese Angaben ließen sich nun dem Wort im Text zuordnen. Suchabfragen könnten sich dann nicht nur auf die Wortformen selbst, sondern auch auf die zu den Wörtern angegebenen Eigenschaften beziehen, sodass es möglich wäre, morphologische sowie teilweise auch stellungsbezogene syntaktische Aspekte in die Suchabfrage einzubeziehen. Der nach der Verknüpfung von Texten und Glossaren noch benötigte manuelle Annotationsaufwand sollte sich damit im günstigsten Fall auf eine bloße Kontrolle der ausgegebenen Daten beschränken können.



gomman-barn *st. n. männliches Kind, masculinum: nom. sg.* 7, 2.
gomo *sw. m. im Compos.* brûtigomo.
got *st. m. deus (dominus): nom.* 1, 1. 4, 14. 5, 9. 13, 14. 21, 7 (3) *etc.* (*zus.* 28 *mal*). got Abrahames (Isakes) 127, 4. got totero 127, 4. truhtin got Israhelo (unser) 4, 14. 128, 2. *voc.* got 118, 2. 3. got min 207, 2 (2). min got 233, 7. *gen.* gotes 82, 9. 90, 4. 126, 3. 244, 2; *vgl.* 4, 18.

Abbildung 1: Auszug aus einem Glossar (Sievers, 1892, S. 343) mit Belegform (z.B. *gotes*) und Belegstelle (z.B. 82, 9) zum Lemma *got*

Dieser Ansatz ist im Forschungsprojekt ‚Referenzkorpus Altdeutsch' auf die althochdeutschen und altsächsischen Textdenkmäler angewendet worden. Hier existiert zu jedem Teilkorpus (mindestens) ein Glossar, das die zur Auszeichnung benötigten Informationen liefern konnte.[1] Während die Texte selbst bereits digitalisiert vorlagen,[2] musste für die Digitalisierung der Glossare erst ein Vorgehen entwickelt werden.

Das DFG-finanzierte Projekt ‚Referenzkorpus Altdeutsch' hat zum Ziel, ein tief annotiertes Korpus aller überlieferten Texte der beiden ältesten Sprachstufen des Deutschen (Althochdeutsch und Altsächsisch, etwa 750 bis 1050 n. Chr.) zu schaffen. Das 650.000 Zeichen umfassende Korpus setzt sich ebenso aus interlinearen Übersetzungen lateinischer Texte wie aus freien Übersetzungen, Adaptationen und gemischten deutsch-lateinischen Texten zusammen. Hinzu kommen einige wenige vollständig in einer altdeutschen Sprache verfasste Texte, vor allem Zaubersprüche.

Im Folgenden wird zunächst das grundsätzliche Vorgehen bei der Digitalisierung beschrieben, bevor anschließend auf die Wiedergabe der Datenstruktur der gedruckten Glossare im Detail eingegangen wird. Den Abschluss bilden ein Abschnitt über spezifische Probleme, die sich beim Digitalisieren ergeben haben, sowie ein Ausblick auf eine mögliche Weiterverarbeitung der digitalen Glossare.

2 Das Vorgehen

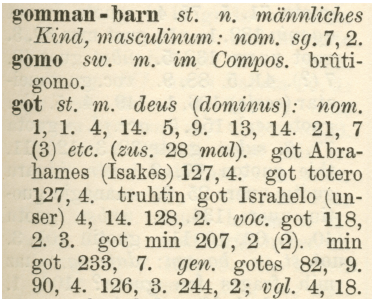
Die Glossare wurden jeweils in ihrer Gänze digitalisiert, da der Aufwand zur Auswahl der tatsächlich benötigten Teile unverhältnismäßig groß gewesen wäre und auch die Gefahr bestanden hätte, Wichtiges auszulassen. Zudem können die digitalisierten Glossare auf diese Weise auch anderen Verwendungen zugeführt werden. Da sich OCR-Programme als ungeeignet für die Drucktypen des späten 19. und frühen 20. Jahrhunderts erwiesen, wurde die Digitalisierung vollständig manuell durchgeführt.[3] Im Gegensatz zu anderen Digitalisierungsprojekten zu Publikationen aus jener Zeit stand hier eine korrekte

Wiedergabe und Unterscheidung verschiedener Typenformen und Drucktechniken in derart hohem Maße im Vordergrund, wie sie automatische Verfahren bislang nicht unter vertretbarem Aufwand erlauben. Zudem war auf diese Weise auch eine unmittelbare Übertragung der erkannten Textteile in eine für die Auszeichnung geeignete Form durchführbar, sodass die Textauszeichnung gemeinsam mit der Digitalisierung geschehen konnte.

Da die Digitalisierung in China erfolgte, war von einer Kenntnis der in den Glossaren verwendeten Sprachen, mit Ausnahme ggf. des Englischen, nicht auszugehen. Die Digitalisierer erhielten jedoch zu jedem Glossar eine bereits digitalisierte und ausgezeichnete Beispielseite als Vorbild, sodass es ihnen möglich war, analog zu verfahren. Im Fall von Ausdrücken, die sich nicht im Druckbild, wohl aber in der Sprache unterschieden, war eine Sprachauszeichnung daher jedoch nicht umsetzbar (vgl. hierzu auch Abschnitt 4). Zugleich verringerte dies aber die Gefahr von Fehlern durch unbewusste Korrektur von Wortformen in Anlehnung an andere Sprachstufen, vor allem an das Neuhochdeutsche.

Um sicherzustellen, dass mit der Digitalisierung zugleich eine möglichst präzise Textauszeichnung erfolgen konnte, war daher vonnöten, den Digitalisierern ein Auszeichnungsschema zur Verfügung zu stellen, das so gut wie möglich auf die Textsorte und die genannten Herausforderungen abgestimmt war und über kurze, aber dennoch in sich eindeutige und klar zuweisbare Tag-Namen verfügte. Denn anders als bei den meisten Wörterbuch-Digitalisierungsprojekten war das Ziel der Digitalisierung nicht die Publikation der Daten, sondern ausschließlich deren interne Weiterverarbeitung, also die Auslesung der Glossardaten, um das zugehörige Textkorpus damit annotieren zu können. Aus diesen Gründen wurde zwar als Auszeichnungssprache der De-facto-Standard XML gewählt, bei der Festlegung eines Schemas (Tag-Sets) allerdings von der Verwendung eines bestehenden Formats – z.B. TEI, wie bei Lemnitzer et al. (im Erscheinen) beschrieben und bei Christmann et al. (2001, S. 25 und passim) angewendet – abgesehen. Stattdessen wurde, ausschließlich für diesen spezifischen Zweck, ein idiosynkratisches Format entwickelt, bei dem insbesondere die Informationen, die später automatisiert ausgelesen werden sollten, zwischen den Glossaren so einheitlich wie möglich dargestellt wurden. Für die Tag-Namen, die im folgenden Kapitel näher beschrieben werden, wurden meist Längen von drei bis fünf Buchstaben, nur bei kombinierten Tag-Namen auch längere Bezeichnungen vergeben. Als Beispiele hierfür seien `<entry>`, `<lem>` (vgl. Abbildung 2) oder `<refLem>` (vgl. Abbildung 8) genannt.[4]

Über die XML-kodierten Beispielseiten hinaus wurde den Digitalisierern zu jedem Glossar je eine Liste der dort verwendeten Elemente, Attribute und Attributwerte sowie eine Liste der Sonderzeichen zur Verfügung gestellt – Letzteres vor allem, um sicherzustellen, dass die Sonderzeichen einheitlich und mithilfe des ihnen entsprechenden Unicode-Characters kodiert wurden. Bei diesem Vorgehen war ausgeschlossen, sämtliche im Glossar auftretenden Fälle im Vorfeld zu erkennen und zu beschreiben, sodass es insbesondere im Falle der ersten bearbeiteten Glossare einer intensiven Korrespondenz mit den Digitalisierern zur Klärung bislang unerfasster Fälle bedurfte.



gomman-barn *st. n. männliches Kind, masculinum: nom. sg.* 7, 2.
gomo *sw. m. im Compos.* brûtigomo.
got *st. m. deus (dominus): nom.* 1, 1. 4, 14. 5, 9. 13, 14. 21, 7 (3) *etc.* (*zus.* 28 *mal*). got Abrahames (Isakes) 127, 4. got totero 127, 4. truhtin got Israhelo (unser) 4, 14. 128, 2. *voc.* got 118, 2. 3. got min 207, 2 (2). min got 233, 7. *gen.* gotes 82, 9. 90, 4. 126, 3. 244, 2; *vgl.* 4, 18.

```
<entry>
  <lem>gomman-barn</lem>
  <pos>st. n.</pos>
  <trlat>männliches Kind,
masculinum</trlat>
  <case>
    <form>nom. sg.</form>
    <inst>
      <rec>7, 2</rec>
    </inst>
  </case>
</entry>
```

Abbildung 2: Beispielhafter Eintrag eines einmal belegten Lemmas (vgl. Sievers, 1892, S. 343)

3 Die Wiedergabe der Datenstruktur

Der strikt hierarchische Aufbau von XML bedingt eine ebensolche Abbildung der Datenstruktur der Glossare. Die oberste Gliederungsebene nach dem Wurzelelement `<root>` bildet meist eine Gliederung nach dem Anfangsbuchstaben des Lemmas (`<let>`, vgl. Abbildung 3). An nächster Stelle in der Hierarchie folgen nun bei allen Glossaren die Lemma-Einträge (`<entry>`). Im Falle der Untergliederung nach Anfangsbuchstaben ist der erste Eintrag innerhalb eines `<let>`-Elements jedoch stets das entsprechende Graphem selbst (`<char>`). Einige Glossare enthalten darüber hinaus noch eine Eigennamenliste, die in der Hierarchie parallel zu den Anfangsbuchstaben in deren Anschluss gestellt wird (`<names>`).

Die Lemma-Einträge sind im Druck stets klar voneinander abgegrenzt und enthalten zunächst das Lemma selbst (`<lem>`), das gegenüber dem übrigen Eintrag deutlich hervorgehoben ist, gelegentlich auch um eine oder mehrere Varianten (`<lemVar>`) davon ergänzt. Anschließend folgen stets Angaben zur Wortart sowie im Falle mancher Wortarten auch zur Flexionsweise (`<pos>`, vgl. Abbildung 2: "`st`[arkes] `n`[eutrum]"). Die selten erfolgende Angabe einer genauen Flexionsklasse kann aufgrund der markierten Darstellung in runden Klammern (z.B., wie in Abbildung 3 gezeigt, bei Hench, 1890) gesondert getaggt werden (`<flex>`).

Meist folgen nun eine oder mehrere mögliche Übersetzungen des Lemmas (`<trlat>`). Die einzelnen Belegfälle (`<inst>`, vgl. Abbildungen 4 und 5[5]) bestehen aus der belegten Form (`<expr>`) – entweder allein oder im Kontext, gelegentlich auch unter Angabe einer zusätzlichen Variante (`<var>`[6]) –, anschließend bei aus dem Lateinischen übersetzten oder übertragenen Texten oft der zugrundeliegenden Entsprechung (`<equi>`) sowie der Belegstelle (`<rec>`), meist bestehend aus Kapitel und Abschnitt, je nach Art des Textes aber etwa auch aus der fortlaufenden Versnummer oder Einzeltextnummer und Zeile in

A

abanst f. (i) invidia: acc. sg. ni uueiz abanst, nescit invidere

```
<root>
  <let>
    <char>A</char>
    <entry>
      <lem>abanst</lem>
      <pos>f.</pos>
      <flex>(i)</flex>
      <trlat l="lat">invidia</trlat>
```

Abbildung 3: Beginn eines Glossars (vgl. Hench, 1890, S. 145)

der Druckausgabe. Zu einer Belegform können auch mehrere Belegstellen genannt sein, an denen diese erscheint; insbesondere dann, wenn bei der Belegform auf eine Angabe des Kontextes verzichtet wird. Wird in einer solchen Aufzählung eine Belegstelle doch gesondert markiert – etwa durch einen erwähnenswerten Kontext oder ein besonderes Äquivalent –, so wird dieser Fall separiert ausgezeichnet (`<subinst>`). Bei flektierbaren Lemmata sind die Belegfälle nach den einzelnen morphologischen Kategorien und ihren Werten (`<form>`) auf bis zu zwei Hierarchieebenen (`<case>`, `<subcase>`) untergliedert. Die kursive Schreibung einzelner Buchstaben innerhalb einer Belegform wird ebenfalls in Form einer gesonderten Auszeichnung übernommen (`<i>`). Auch wenn zu einem im Kontext angegebenen Beleg weitere vergleichbare Belegstellen genannt sind, wird dies markiert (`<sim>`, vgl. Abbildung 7). Fußnoten zu einer Seite werden sowohl einzeln (`<fn>`) als auch im Block (`<foot>`) getaggt.

Einige Lemmata weisen zudem eine teilweise sehr differenzierte semantische Gliederung auf (`<usage>`, `<subusage>`, `<specusage>`, `<subspecusage>`, vgl. Abbildung 6). Oberhalb davon erscheinen in manchen Fällen noch bis zu zwei weitere Kategorien zu unterschiedlichen Wortarten oder etwa unterschiedlichen folgenden Kasus bei Präpositionen (`<qual>`, `<nat>`[7]). Die Zuordnung geschieht hier analog zur Auszeichnung im Glossar: Bei Sehrt (1966) etwa werden mit `<usage>` immer die arabischen Zahlen in der Gliederung ausgezeichnet, unabhängig davon, ob bei einzelnen Lemmata darüber hinaus auch `<qual>` (hier Großbuchstaben) oder `<nat>` (hier römische Zahlen) vorkommen. Wie in Abbildung 6 dargestellt, erscheinen zuweilen auch dem Lemma untergeordnete Komposita (`<comp>`) hier als semantische Unterkategorie eingruppiert.

Über das bisher Beschriebene hinaus sind bei Sehrt (1966) innerhalb eines Lemmas semantische und morphologische Gliederung voneinander getrennt: Zunächst werden charakteristische Verwendungsweisen des Lemmas im Kontext dargestellt, anschließend folgen die bloßen belegten Formen, angeordnet nach Flexionskategorien (vgl. Abbildung 7). Da große Teile des *Heliand* eine parallele Überlieferung in mehreren Handschriften

hêr adj. exalted: comp. nom. sg. m. subst. *herro*, dominus 12, 21, — superl. hêrôsto, princeps: gen. sg. m. herostin 21, 21,

```
<usage>
  <case>
    <form>comp. nom. sg. m. subst.</form>
    <inst>
      <expr>h<i>erro</i></expr>
      <equi l="lat">dominus</equi>
      <rec>12, 21</rec>
    </inst>
  </case>
</usage>
<usage>
  <case>
    <form>superl.</form>
    <expr>hêrôsto</expr>
    <subcase>
      <form>gen. sg. m.</form>
      <inst>
        <expr>herostin</expr>
        <rec>21, 21</rec>
      </inst>
```

Abbildung 4: Morphologische Gliederung bei Hench (1890, S. 170)

mit diu 137, 2. 144, 1. 145, 1. 151, 4 (*dum*). 178, 4. 210, 5;

```
<inst>
  <expr>mit diu</expr>
  <rec>137, 2</rec>
  <rec>144, 1</rec>
  <rec>145, 1</rec>
  <subinst>
    <equi>dum</equi>
    <rec>151, 4</rec>
  </subinst>
  <rec>178, 4</rec>
  <rec>210, 5</rec>
</inst>
```

Abbildung 5: Darstellung gleichlautender Belege bei Sievers (1892, S. 464)

an (*got.* ana, *an.* á, *ahd.* ana an, *ags.* on, *afries.* ana an *FT 11*) *A adv.* (*Syn. § 10*) *an, auf, nach*; *hinan, hinauf*: *1) in Verbindung mit einem Verbum und dat. pers.* (*Syn. S. 211, 215, 217*): an uuas imu anst godes 784. that he themu uuîbe gedorsti stên an uuerpen 3877. dedun im eft ôđer (lakan) an 5498. than ûs liudi farad an 4141. *1a*) *c. adv. et verb.* (*vgl. Germ. XI, 214*) than hêr theobas an thingstedi halden 3745. *2*) *c. verb. et acc. pers.* (*Syn. S. 207, 214*): sah sie an langо 1291. thô bigan ina Crist sehan an mid is ôgun 3281. that sie ina than feteros an leggien môstin 3796. uueldun ina the andsacon stên an (ana) uuerpan 3941; 3871, 3946. that mugun uui ina gitellian an 5189. thes (*für* that) sie an iro môd spenit 1354. *3*) *mit einem Verbum ohne persönl. Objekt* (*Syn. S. 89, 207, 215*): huuat gi sculun an hebbean uueros te gewêdea 1664. bûtan sô gi than an hebbean te gareuuea 1856. bigunnun im (*reflex.*) tellien an 5072. *4*) *mit einem Verbum und adverb. Ausdrucke verbunden* (*Syn. §§ 168, 339*): that sie môstin an faren an thiu berhtun bû 3653.

an *1*) **784 3877. 4141**, 5498*; *1a*) **3745**; *2*) **1291** (*V*). **1354** (*V*), **3281, 3796, 5189**; *3*) **1664, 1856. 5072**; *4*) **3653.** *an 1516 *M.* ana (an*) *2*) 3871 *M* (*C* auuerpe), **3941, 3946.** *B. präp.* (*Syn. §§ 163, 165, 237*) *I c. dat.*[1]) *1) rein räumlich, in* (*unter*), *an, auf, bei*: *a) zu einem Verbum tretend um den Ort zu bezeichnen, wo das durch das Verbum Ausgedrückte stattfindet*: *α) bei Verbis, welche an sich eine Ortsbeziehung voraussetzen oder eine Ergänzung durch eine Ortsbestimmung fordern, wie bei den intransitiven Verbis des sich Befindens, Verweilens, Vorhandenseins, Existierens*: thea liudi the hêr nu langо bidun an thesara middilgard 524. ne ik an them bendion mid thi bîdan uuillie 4682; 4947. bûan an them burugium *Gen.* 238. the habda an (*M* at) them uuîha sô filu uuintro endi sumaro gilibd an them liohta 465—66; *Gen.* 92. libbian an thesun lande *Gen.* 71, 76, 305, 333. *übertragen*: that ik scal an thînun heti libbian, forđ an thînun fîundscepi *Gen.* 60—61. ligid that kind an ênera cribbiun 407; 2140—41, 3364. huuô hîr uuegos tuêna liggead an thesumu liohte 1772; 1782. (ik) an feteron lag biklemmid an karkare 4399—4400; 5397. an is breostun lag 4602. liggian an ênam diapun dala *Gen.* 29. that corn, that thar an theru lêian gilag[2]) 2394. môste thar thô an thes mahtiges Kristes barme restien 4601; 2134—34—35. sâton ira heritogon an landо gihuem 59. sitit an is uuînseli 229; 549. thar he an is rîkea sat 716. thar sie an (*M* at) mahle sittiad 1312. thar he an is benki sat 2746; 5269. Lazarus sat bliđi an is barme 3362. sittian an them stênuuege 5462. he an themu uuîhe stôd 4240. Hiericho, thiu thar an Iudeon stâd 3625. he an middien stôd 3908. stêt thînes brôđor uurâca bitter an helli *Gen.* 79. thiu burg, thiu an berge stâd 1395. thie uurti, thea hîr an felde stâd 1673; 1680. the ubilo bôm, thar he an erđu stâd 1745. thuo

```
<entry>
  <lem>an</lem>
  [...]
  <qual n="B">
    <pos>präp.</pos>
    <ref>Syn. §§ 163, 165, 237</ref>
    <nat n="I">
      <pos r="13_1">c. dat.</pos>
      <usage n="1">
        <trlat l="deu">rein räumlich, in (unter), an, auf, bei</trlat>
        <subusage n="a">
          <trlat l="deu">zu einem Verbum tretend um den Ort zu
bezeichnen, wo das durch das Verbum Ausgedrückte stattfindet</trlat>
          <specusage n="α">
            <trlat l="deu">bei Verbis, welche an sich eine Ortsbeziehung
voraussetzen oder eine Ergänzung durch eine Ortsbestimmung fordern,
wie bei den intransitiven Verbis des sich Befindens, Verweilens,
Vorhandenseins, Existierens</trlat>
             [...]
             <subspecusage>
               <comp>an-innan</comp>
               <comp>an-uppan</comp>
```

Abbildung 6: Semantische Gliederung mit sechs Hierarchieebenen (gekürzt) bei Sehrt (1966, S. 13 f.)

aufweisen, wird hier auch dargestellt, welche Form an welcher Stelle in welcher Handschrift erscheint und welche Unterschiede zwischen den Schreibungen der einzelnen Manuskripte vorliegen. Hierbei werden die sonst als Belegfälle behandelten Beispiele innerhalb der semantischen Gliederung anderweitig ausgezeichnet (`<ex>`), während die Darstellung der Belegformen im Kontext selbst unverändert bleibt (`<expr>`), sodass die Auszeichnung durch `<inst>` den tatsächlichen Belegfällen innerhalb der morphologischen Gliederung vorbehalten wird. Hier werden nun allerdings die belegten Formen anders markiert (`<shape>`), da sie zwar ohne Kontext, aber unter Angabe des Manuskripts und Bezugnahme auf die semantische Gliederung aufgeführt sind.

Schließlich können innerhalb der Lemmata auch erläuternde Teilübersetzungen von Ausdrücken oder ihren lateinischen Entsprechungen (`<expl>`) sowie Anmerkungen (`<rem>`) und Bezüge bzw. Verweise (`<ref>`) auftreten (vgl. Abbildung 8). Vor allem Letztere enthalten Bezugnahmen (`<refLem>`) oder Verweise (`<refEntry>`) auf andere Lemmata sowie Bezüge auf Belegformen (`<refForm>`) oder Ausdrücke (`<refExpr>`), oft eingeleitet durch ein `<cf>cf.</cf>` oder `<cf>s.</cf>`. Auch die nicht anderweitig markierten Inhalte einer Anmerkung werden gesondert als Kommentare ausgezeichnet

firin-uuord *stn.* (*Syn. S. 9*) *Frevelwort, Schmähung*: felgidun imu firinuuord [9]), bismersprâka 5116; 5299.

acc. pl. firinuuord 5116, 5299*.

firio *siehe* firihos.

firi-wit (*ahd.* firi-wizzi, *ags.* fyr-wit *FT 231, 410*) *stn.* (*Syn. S. 10, 64,*

nm. zu 2428. 3) *ibid. S. 450, 32—33.*
S. 430, 13.
) *ibid. S. 465, 7; 398, 14.*
III, 203. 9) *Sievers, S. 430, 13.*

```
<entry>
  <lem>firin-uuord</lem>
  <pos>stn.</pos>
  <ref>Syn. S. 9</ref>
  <trlat l="deu">Frevelwort, Schmähung</trlat>
  <ex>
    <expr r="133_9">felgidun imu firinuuord, bismersprâka</expr>
    <rec>5116</rec>
    <sim>
      <rec>5299</rec>
    </sim>
  </ex>
  <case>
    <form>acc. pl.</form>
    <inst>
      <shape m="M, C">firinuuord</shape>
      <rec>5116</rec>
    </inst>
    <inst>
      <shape m="C">firinuuord</shape>
      <rec>5299</rec>
    </inst>
  </case>
  <foot>
    <fn n="133_9">Sievers, S. 430, 13</fn>
  </foot>
</entry>
```

Abbildung 7: Gliederung nach Handschriften bei Sehrt (1966, S. 133), vollständiger Lemma-Eintrag

oder an welchem sie stattfindet. Hinter githuagan ist *zu interpunktiren und* reino *als 3. Ps. sing. Conj. des Verbums* reinôn *aufzufassen, wird eben durch Joh. 13, 10, worauf man sich beruft, zurückgewiesen. Dort heisst es nämlich:*

```
<rem>
  <com>Hinter</com>
  <expr>githuagan ist</expr>
  <com>zu interpunktiren und</com>
  <refForm>reino</refForm>
  <com>als 3. Ps. sing. Conj. des
Verbums</com>
  <refLem>reinôn</refLem>
  <com>aufzufassen, wird eben
durch</com>
  <bib>Joh. 13, 10</bib>
  <com>, worauf man sich beruft,
zurückgewiesen.</com>
</rem>
```

Abbildung 8: Beispiel einer Anmerkung mit diversen Referenzen (vgl. Kelle, 1881, S. 156 f.)

(`<com>`). Daneben werden auch Verweise auf Bibelstellen markiert (`<bib>`).

Verschiedene weitere Informationen werden schließlich in Form von Attributen umgesetzt – etwa die Aussage darüber, ob über eine Angabe Zweifel bestehen (z.B. `d="?"`), die Angabe der Sprache (z.B. `l="lat"`), die einheitlich gemäß ISO 639-3[8] erfolgt, die Angabe des der angegebenen Form zugrundeliegenden Manuskripts (z.B. `m="C"`, vgl. Abbildung 7) und die Angabe einer Referenz (z.B. `r="133_9"`). Auch sonstige nummernartige Bezeichnungen, z.B. von Fußnoten oder semantischen Hierarchieebenen, werden markiert (z.B. `n="1"`).

4 Besondere Probleme der Digitalisierung

Bereits angesprochen wurde der Aspekt, dass nicht alle Problemfälle auf einer Beispielseite behandelt werden können (s. Abschnitt 2), was insbesondere bei den ersten digitalisierten Glossaren zu einer weiteren Überarbeitung und Erweiterung des Tag-Sets führte. Und auch wenn ein Großteil der Elemente sich in allen Glossaren einsetzen ließ, sind einige von ihnen nur für ein einziges Glossar konzipiert worden, wie in Abschnitt 3 exemplarisch dargestellt.

Der Umstand, dass XML eine strikte hierarchische Elementstruktur aufweist, erwies sich im Falle sich überschneidender Hierarchien als großer Nachteil. Ein eindrucksvolles Beispiel hierfür bieten die Fußnoten. Diese sind der Ebene der Seite untergeordnet, die für die Struktur des Glossars an sich aber keine Rolle spielt. Eine Lösung hierfür könnte sein, die Fußnoten umzunummerieren, indem man sie in eine vom übrigen Inhalt separierte Ebene am Ende des Dokuments überführt. Als weniger umständlich hat sich jedoch erwiesen, die Fußnoten parallel zu den (vollständigen) Lemma-Einträgen

anzuordnen und dazu an das Ende des Lemma-Eintrags, auf den sie sich beziehen, zu verschieben. In beiden Fällen ist es allerdings nötig, zur Fußnote die Seitenzahl dazuzumarkieren, um eine eindeutige Zuordnung zu gewährleisten.

Die Verarbeitung von Varianten wird oft dadurch erschwert, dass diese nur abgekürzt dargestellt werden – so handelt es sich bei dem in Fußnote 6 dargestellten Beispiel `<expr>ahtu</expr><var>ahto</var>` um eine Digitalisierung von gedrucktem „ahtu (*var.* -o)“ (vgl. Sievers, 1892, S. 302). Entscheidend ist jedoch, diese in ihrer vollständigen Form zu digitalisieren, um sie so später automatisch auslesen zu können. Dies stellt oft ein besonderes Problem für die Digitalisierer dar. Bei der Darstellung von Alternativen – etwa Lemmaformen, Flexionskategorien oder belegten Formen – kann zudem keine Art der Wiedergabe in XML völlig verhindern, dass bei der Vorbereitung des maschinellen Auslesens der Daten auf ihre Berücksichtigung besonders geachtet werden muss, um sie nicht zu übergehen. Während innerhalb eines Lemma-Eintrags angegebene Komposita, die an anderer Stelle separat erscheinen, bloß als solche markiert zu werden brauchen, stellt sich die Frage, inwieweit Sublemmata eigene Einträge konstituieren sollten. Mehr-Wort-Lemmata bereiten bei der XML-Digitalisierung zwar keine Schwierigkeiten, werden allerdings aufgrund ihrer Seltenheit beim Auslesen so eventuell nicht erwartet und könnten nicht erfasst werden – wenn in die Suchabfrage etwa keine Leerzeichen einbezogen werden.

Trotz der Bereitstellung von Listen der vorkommenden Sonderzeichen können auch bei der Kodierung von Buchstaben Fehler auftreten, am häufigsten durch Fehlinterpretation aufgrund von Veränderungen der Glyphe beim oder seit dem Druck oder Darstellung in mangelhafter Qualität nach dem Einscannen der Bücher für die Digitalisierer (z.B. `c` für ‹e›, `r` für ‹n› oder auch `n` für ‹u›). Selten erscheinende Sonderzeichen, die nicht in der Sonderzeichenliste enthalten sind, können zudem fehlinterpretiert und falsch kodiert werden. Dies kann etwa bei Angabe einer Lemmaentsprechung in einer außergermanischen Sprache oder bei ungewöhnlichen Beleg-Schreibungen (z.B. Wiedergabe von belegtem ‹ŏ› als `ô`) vorkommen. Hinzu kommt das Problem von aus der Printausgabe übernommenen Druckfehlern. Wenngleich einzelne Fehler systematisch auftreten und daher mithilfe von Ersetzungsregeln automatisiert korrigiert werden können, hat sich die Datenqualität der digitalisierten Glossare insgesamt als so gut erweisen, dass sich für den Zweck des Projekts eine systematische Fehlersuche erübrigte und Korrekturen meist nur einzelfallbezogen erfolgen, wenn bei der Weiterverarbeitung der Daten ein Fehler offensichtlich wird.

Wie die Abweichungen im Druck sind auch semantische Unterschiede, die nicht explizit markiert sind, für Digitalisierer, die die im Glossar verwendeten Sprachen nicht beherrschen, nicht zu erkennen. Dies gilt insbesondere etwa für Lemmaübersetzungen, deren Sprache aufgrund der für einen Philologen des 19. bzw. 20. Jahrhunderts bestehenden Offensichtlichkeit oft nicht angegeben ist. Im Falle der Glossare zum Althochdeutschen und Altsächsischen handelte es sich dabei um Deutsch, Englisch und Latein, auf deren Angabe daher bei der Digitalisierung verzichtet werden musste, sofern die Position oder Darstellung von Ausdrücken in den einzelnen Sprachen die Sprache nicht eindeutig erkennen ließ (vgl. Abbildung 2).

Auch die Abbildung der semantischen Strukturierung – einschließlich der Gliederung eines Lemmas in mehrere Wortarten – mit bis zu sechs Hierarchieebenen innerhalb eines Lemmas (vgl. Abbildung 6) verlangt gerade bei sehr häufigen Lemmata genaue Abstufungen; hier kommen nicht selten auch Fehler im gedruckten Glossar vor.

Innerhalb von Erläuterungen und Kommentaren erscheinen bisweilen interne sowie externe Querverweise auf andere Lemmata, Sublemmata oder Wortformen und Mehrwortausdrücke samt Übersetzung und Textstellenangabe. Diese sind ebenso wie Kommentare an nahezu jeder denkbaren Stelle innerhalb des Lemma-Eintrags – etwa mitten in Mehrwortausdrücken, die einen Beleg enthalten – unter Wahrung der hierarchischen Struktur sowie der Zusammengehörigkeit ihres Kontextes entsprechend auszuzeichnen.

5 Zur Verwendung der Daten

Der Einsatz der digitalisierten Glossare im Textkorpus wird bei Linde und Mittmann (im Erscheinen) ausführlich dargestellt. Hier wird auch auf die Grenzen der Glossardaten und ihrer Verknüpfbarkeit mit dem Text eingegangen und erläutert, wie Lemma-Formen und Übersetzungen aus den Glossaren an einen Standard angepasst werden, der für alle Texte der jeweiligen Sprachstufe Verwendung finden kann. Diese Standardisierung wäre zudem auch die Voraussetzung für die Aufnahme der Glossare in ein digitales Wörterbuchnetz, wie es Burch und Rapp (2007) beschreiben.

6 Anhang: Ein Anwendungsbeispiel

Abbildung 9 zeigt eine mit Hilfe von Glossar-Daten automatisiert vorgenommene Vorannotation der Wortform *gommanbarn* im althochdeutschen *Tatian* (vgl. Abbildung 2 sowie Sievers, 1892, S. 25 und 343; leicht vereinfachte Darstellung). Die obersten beiden Zeilen enthalten das annotierte Wort, die letzten beiden Zeilen geben dessen Position im Text an. In den dazwischen befindlichen Zeilen erscheint die Lemmatisierung auf Grundlage des Glossars sowie der dort angegebenen weiteren Informationen, bereits umgewandelt in ein standardisiertes Format.[9] In einem nächsten Schritt müssen die Angaben nun manuell geprüft werden, bevor die Annotation in die Datenbank überführt werden kann. Weitergehende Erläuterungen zum genauen Vorgehen und zu dabei auftretenden Problemfällen finden sich bei Linde und Mittmann (im Erscheinen) sowie Linde (in diesem Band).

Anmerkungen

[1] Für das Althochdeutsche sind Heffner (1961), Hench (1890, 1893), Kelle (1881), Sehrt (1955) sowie Sievers (1874, 1892) verwendet worden, für das Altsächsische Sehrt (1966) und Wadstein (1899).

[2] Die Digitalisierung erfolgte im Rahmen des TITUS-Projekts (Thesaurus Indogermanischer Texte und Sprachen) – `http://titus.uni-frankfurt.de` –, über diese Seite sind auch die Texte abrufbar.

[3] Für ein vergleichbares Vorgehen vgl. Christmann et al. (2001, S. 23).

[4] Da die Glossare vollständig erfasst wurden, wäre eine spätere Überführung des idiosynkratischen XML-Formats in ein standardisiertes XML-Format (etwa TEI), um eine Publikation der Glossare zu ermöglichen, prinzipiell denkbar. Hierzu müssten die hierarchischen Strukturen angepasst und

Referenztext Wort	gommanbarn									
Referenztext Buchstaben	g	o	m	m	a	n	b	a	r	n
Lemma	gommanbarn									
Übersetzung	männlicher Nachkomme									
Wortart Lemma	NA									
Wortart Beleg	NA									
Flexion Lemma	a,z_Neut									
Flexion Beleg 1	a,z_Neut									
Flexion Beleg 2	Sg_Nom									
Kapitel	7									
Unterkapitel	2									

Abbildung 9: Automatisierte Vorannotation eines Wortes auf Grundlage von Glossardaten (NA: Nomen, Appellativum; a,z: germanische Nominalflexionsklassen)

die Namen der Elemente, Attribute sowie Attributwerte in den entsprechenden Standard überführt werden (vgl. hierzu und zum Folgenden die Abschnitte 3 und 4). Dabei müssten insbesondere die ungewöhnlicheren Merkmale der Glossare – die Angabe von Belegstellen, die Nennung von Belegformen im Kontext ohne Markierung der Belegform selbst, die Einfügung von Kommentaren, internen und externen Referenzen an nahezu jeder denkbaren Stelle sowie das Vorkommen von Druckfehlern – berücksichtigt werden. Zudem ist die digitalisierte Form der Glossare trotz der hohen Datenqualität nicht systematisch auf verbleibende Druck- oder neu hinzugekommene Lesefehler geprüft worden. Auch dass bei Übersetzungen oftmals die Angabe der Sprache fehlt, würde eine Nutzung dieser Angaben erschweren. Eine standardisierte digitale Form der Glossare herzustellen, wäre unter Berücksichtigung dieser kleineren Erschwernisse (von denen einige jedoch auch bei standardkonformer Digitalisierung bestanden hätten) möglich, erwies sich jedoch, wie angeführt, für das Projektvorhaben als nicht optimal.

[5]Verweise auf Abbildungen gelten in diesem Artikel i. d. R. auch für die im Folgenden genannten Tags, solange kein anderweitiger Verweis erfolgt.

[6]Z.B. `<expr>ahtu</expr><var>ahto</var>` zum Lemma *ahtu* 'acht' bei Sievers (1892, S. 302).

[7]Die Benennung erfolgte durch Abkürzung der englischen Termini *quality* und *nature*, jeweils im Sinne von 'Beschaffenheit, Eigenschaft'.

[8]Nähere Informationen zu der Norm finden sich bei der zuständigen Registrierungsstelle SIL International unter `http://www.sil.org/iso639-3/`.

[9]Die Übersetzung ist ebenfalls bereits standardisiert. Da eine Belegform im Glossar nicht explizit angegeben ist, wird die Lemmaform hier als Belegform angenommen. Der in Abbildung 2 dargestellte Bindestrich ist hier regelmäßig getilgt, zumal er nicht in den Belegformen erscheint und diese so besser mit den Lemmata verglichen werden können.

Literatur

Burch, T. & Rapp, A. (2007). Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven. In: D. Burckhardt, R. Hohls & C. Prinz (Hrsg.), *Beiträge der Tagung .hist 2006 = Historisches Forum* (Bd. 10, S. 607–627). Berlin: Clio-online.

Christmann, R., Hildenbrandt, V. & Schares, T. (2001). Ein "heiligthum der sprache" digitalisiert: Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm auf CD-ROM und im Internet. In: N. Castrillo Benito & P. Stahl (Hrsg.), *TUSTEP educa. Actas de Congreso del Interna-*

tional TUSTEP User Group. Peñaranda de Duero (Burgos) Octubre 1999 (S. 13–37). Burgos. (http://kompetenzzentrum.uni-trier.de/files/8513/1349/4217/Grimmbu.pdf)

Heffner, R.-M. S. (1961). *A Word-Index to the Texts of Steinmeyer. Die kleineren althochdeutschen Sprachdenkmäler.* Madison: The University of Wisconsin Press.

Hench, G. A. (1890). *The Monsee Fragments.* Straßburg: Trübner.

Hench, G. A. (1893). *Der althochdeutsche Isidor.* Straßburg: Trübner.

Kelle, J. (1881). *Glossar der Sprache Otfrids.* Regensburg: Manz.

Lemnitzer, L., Romary, L. & Witt, A. (im Erscheinen). Representing human and machine dictionaries in Markup languages (SGML, XML). In: R. H. Gouws, U. Heid, W. Schweickhard & H. Erns (Hrsg.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography. Handbooks of Linguistics and Communication Science (HSK).* Berlin/New York: de Gruyter. (http://www.dwds.de/media/publications/text/Lemnitzer_Romary_Witt-HSK-Article-v2009-12-15_-_deformatted.pdf)

Linde, S. & Mittmann, R. (im Erscheinen). Old German Reference Corpus. Digitizing the knowledge of the 19th century. In: P. Bennett, M. Durrell, S. Scheible & R. J. Whitt (Hrsg.), *New Methods in Historical Corpus Linguistics = Korpuslinguistik und interdiziplinäre Perspektiven auf Sprache – Corpus linguistics and Interdisciplinary perspectives on language (CLIP)* (Bd. 3). Tübingen: Narr.

Sehrt, E. (1955). *Notker-Wortschatz.* Halle: Niemeyer.

Sehrt, E. (1966). *Vollständiges Wörterbuch zum Heliand und zur altsächsischen Genesis* (2. Aufl.). Göttingen: Vandenhoeck & Ruprecht.

Sievers, E. (1874). *Die Murbacher Hymnen.* Halle: Buchhandlung des Waisenhauses.

Sievers, E. (1892). *Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar* (2. Aufl.). Paderborn: Schöningh.

Wadstein, E. (1899). *Kleinere altsächsische Sprachdenkmäler.* Norden/Leipzig: Soltau.

Sonja Linde

Manuelle Abgleichung bei automatisierter Vorannotation:
Das Tagging grammatischer Kategorien im *Referenzkorpus Altdeutsch*

Abstract

Das Referenzkorpus Altdeutsch überführt die grammatischen Informationen der etablierten Textwörterbücher zu den ältesten deutschen Texten in ein Annotationstool, wobei die so automatisiert gewonnenen Daten für jede einzelne Wortform manuell überprüft und angeglichen werden müssen. In diesem Artikel werden verschiedene Fälle vorgestellt, in denen die grammatische Annotation des Referenzkorpus Altdeutsch von den aus den Grammatiken und Glossaren implementierten Angaben abweicht. Hierbei handelt es sich vor allem a) um Alternativdeutungen zu den traditionellen Interpretationen bestimmter Formen, b) um Lücken in der Beschreibung bestimmter Formen, die aus einem ausschließlich historisch ausgerichteten Grammatikverständnis entstehen und c) um Fehler, die aufgrund einer nicht eindeutigen Beschreibung entstehen.

1 Einleitung

Das Referenzkorpus Altdeutsch[1] digitalisiert und annotiert sämtliche Texte der althochdeutschen und altsächsischen Überlieferung (ca. 750 – 1050 n. Chr.). Die Annotation umfasst linguistische und strukturelle Informationen wie auch die Metadaten eines Textes.

Zur Bearbeitung werden die möglichst handschriftengetreuesten Textausgaben in das Annotationstool Elan[2] überführt und die jeweiligen Textwörterbücher und Glossare[3] implementiert. Die so automatisiert gewonnenen lemmabezogenen grammatischen Informationen müssen für jede einzelne Wortform manuell überprüft und dem Kontext entsprechend angeglichen werden. Die präzisen, umfangreichen und mit ungemein großer Sachkenntnis zusammengetragenen Angaben der zum Großteil noch aus dem 19. Jhd. stammenden Glossare, die für alle größeren altdeutschen Texte verfügbar sind, bieten dem Korpus somit eine wertvolle Grundlage.

Für die Auszeichnung der PoS-Informationen und der morphologischen Werte unterscheidet das Referenzkorpus Altdeutsch zwischen einer lemma- und einer belegbezogenen Kennzeichnung. Während auf der PoS-Lemmaebene (L) grundsätzlich nur die Werte angegeben werden, die im Lemma selbst festgelegt sind – also die Grundmarkierungen der Wortarten bzw. die unveränderlichen, im Lexikon verzeichneten morphologischen Informationen – werden auf der PoS-Belegebene (B) diese Werte spezifiziert. Prinzipiell werden dabei für die Lemmaebene (L) die in den einzelnen Wörterbüchern aufgeführten Informationen zur Wortart übernommen. Lediglich bei wenigen Ausnahmen werden diese angeglichen, z.B. bei abweichenden grammatischen Annahmen des Referenzkorpus Altdeutsch zu den Informationen der Glossare und Wörterbücher (vgl. hierzu Abschnitt 3).

2 Manuelle Angleichung bei automatisierter Vorannotation

Zur Differenzierung zwischen Lemma- und Belegebene werden die PoS-Kategorien hinsichtlich ihrer Hauptwortart und ihrer konkreten syntaktischen Verwendungsweise unterschieden.[4]

(1)	Wortform	*lag*	*liggian*	*liggiandi*	*liggeandan*
	Lemma	*liggian*	*liggian*	*liggian*	*liggian*
	PoS L	VV	VV	VV	VV
	PoS B	VVFIN	VVINF	VVPS	VVPSA

Bei der morphologischen Beschreibung ist die Kennzeichnung der Flexionsklasse des Belegs dort präziser, wo die Flexionsklasse des Lemmas nicht eindeutig ist, der Beleg aber eine Festlegung erlaubt. So flektiert z.B. das ahd. Lemma *man* ‚Mann' ursprünglich konsonantisch, in den Singularformen erscheint kein Flexionsmorphem (NGDA *man-Ø*). Diesem Lemma wird entsprechend für die konsonantische Deklination der Wert C zugewiesen. Daneben erscheinen aber bereits zu althochdeutscher Zeit Formen der maskulinen a-Stämme (G *mannes*, D *manne)* und in diesen Fällen erhält *man* den Wert a für die Deklinationsklasse. Da *man* im Ahd. also sowohl konsonantisch als auch analog zu den a-Stämmen flektiert, erhält das ahd. Lemma die Werte beider Flexionsklassen (Fl(exions)klasse L(emma)). Erst bei eindeutigen Wortformen wird die Angabe der Flexionsklasse präzisiert (Fl(exions)klasse B(eleg)).

(2) *Yrhugis thar thoh éines man* (O 2,42)
gedenkst da doch eines Mann-SG.GEN
„(Du) gedenkst da doch eines Mannes"

Lemma	*man*
PoS L, B	NA
Fl.klasse L	C,a_Masc
Fl.klasse B	C_Masc

(3) *inti ther mannes sun uuirdit giselit* (T 153, 2)
und der Mann-SG.GEN Sohn wird verraten
„Und der Menschensohn wird verraten werden"

Lemma	*man*
PoS	NA
Fl.klasse L	C,a_Masc
Fl.klasse B	a_Masc

Durch die Differenzierung zwischen Lemma und Beleg bestehen Möglichkeiten von statistischen Abgleichen bestimmter Verteilungen wie z.B. von Vorkommen von Deklinationsklassen eines Subkorpus. Das Lemma *man* ‚Mann, Mensch' flektiert ursprünglich konsonantisch und tritt in althochdeutscher Zeit zunehmend stark flektierend nach dem Vorbild der

maskulinen a-Stämme auf, weswegen beide Flexionsformen im Wörterbuch verzeichnet sind und entsprechend für die Flexionsklasse des Lemmas (Flexion L) aufgeführt sind.

Im Tatian tritt das Lemma *man* ‚Mann, Mensch' 266mal auf. Davon werden 157 Belege sowohl als konsonantisch als auch als a-Stamm beschrieben, da es sich hier um formgleiche Kasus handelt. Die übrigen Belege können aufgrund ihres eindeutigen Flexionsmorphems einer Deklinationsklasse zugewiesen werden, die a-Stämme (70mal) treten dabei fast doppelt so häufig wie die konsonantischen Formen (39mal) auf. Die folgende Tabelle zeigt die Belegzahlen für einzelne Kasusformen von *man* im Tatian.

Lemma / Form	C, a	C	a	gesamt
man	157	39	70	266
Sg. Nom	75	-	-	75
Gen	-	-	57	57
Dat	-	6	13	19
Akk	29	-	-	29
Pl. Nom	-	28	-	28
Gen	19	-	-	19
Akk	-	5	-	5
Dat	34	-	-	34

Tabelle 1: Vorkommen von *man* im Tatian nach Flexionsklasse und Form

Neben der Spezifizierung der von den Wörterbüchern dem Lemma zugewiesenen Werte dient die Unterscheidung zwischen Lemma- und Belegebene auch der Beschreibung von Sprachwandelprozessen.

Im Ahd. und As. gibt es für den Plural und das Femininum kein Possessivpronomen, in diesen Fällen wird der jeweilige Genitiv des Personalpronomens (Fem. Sg. *ira*, Pl. *iro*) verwendet. Durch die Trennung der PoS-Kennzeichnung in Lemma und Beleg kann so z.B. die kontextbezogene Verwendung von ahd./as. *iro* ‚ihr' als nicht flektierendes Possessivpronomen mit dem Tag DPOS auf der Ebene PoS B(eleg) dargestellt werden. Die lemmabezogene formale Bedeutung als Genitiv Plural des Personalpronomens wird mit dem Tag PPER auf der Ebene PoS L(emma) und durch die Angabe der veränderlichen flexivischen Werte auf einer morphologischen Ebene als Genitivform beschrieben.

(4) *After thiu her uuosc iro fuozzi.* (T 156,1)
danach er wusch er-MASC.GEN.PL. Füße-PL.ACC.
„Danach wusch er ihre Füße" (eigentlich: die Füße ihrer = die Füße von ihnen)

Lemma	ër
PoS L	PPER
PoS B	DPOS
Flexion	*Masc.Gen.Pl.3*

Bei der Annotation der meisten Wortformen folgt das Referenzkorpus Altdeutsch weitestgehend den Angaben der Wörterbücher, d.h. die im Annotationstool für die jeweilige Ebene angegebenen Werte entsprechen den Angaben in den Glossaren und auf der Belegebene werden von den Annotatoren ggf. Entscheidungen zugunsten des einen oder anderen Wertes getroffen.

Überdies erscheinen aber Fälle, bei denen das Referenzkorpus Altdeutsch von den Beschreibungen der Textwörterbücher und Grammatiken[5] abweicht.

Zum einen kann es sein, dass das reichhaltige Wissen der Wörterbücher und der Grammatiken nicht mehr dem aktuellen Vorgehen bei der Beschreibung von grammatischen Phänomenen entspricht, und deswegen eine alternative Kennzeichnung entwickelt werden muss. Teilweise treten systematisch Lücken in der Beschreibung auf, zumeist, weil entweder das Wissen von den Verfassern vorausgesetzt wurde oder weil bestimmte Unterschiede nicht erfasst wurden. In einigen (wenn auch wenigen) Fällen führt die durchgehende Systematisierung im Sinne einer Vereinfachung von verschiedenen Beschreibungsmöglichkeiten bei durchweg komplexer Beleglage zu fehlerhaften Vorannotationen.

Mit anderen Worten: In welchen Fällen stößt die Nutzung der altdeutschen Wörterbücher und Grammatiken trotz deren Kenntnisreichtums, deren Detailgenauigkeit und Fülle an Belegen für ein computerbasiertes, modernen Ansprüchen genügendes Korpus an ihre Grenzen?

3 Konflikte zwischen traditioneller Grammatikbeschreibung und moderner Theorie: Alternativentscheidungen

Bereits in den ältesten Texten des Deutschen treten mehrgliedrige Konjunktionen und Adverbien auf.

Die Instrumentalis-Form *thiu* des Demonstrativpronomens ahd. *ther*, as. *the* ‚der' erscheint oftmals in Verbindung mit einer Adposition oder einem Adverb. Diese Zusammensetzungen mit *thiu* haben häufig zwei Lesarten und treten dann sowohl als Adverb bzw. Präpositionaladverb als auch als Subjunktion auf.

(5) *Than mag man dragan* ADV[*aftar thiu*] *lîhtlîcora lîđ*
dann kann man-PI auftragen danach leicht-COMP Wein
(H 2054)
„dann soll man danach den leichteren Wein auftragen"

(6) SUBJ[*After thiu*] *sie inan erhiengun, intfiengun sin giuuati (...)*
nachdem sie ihn erhängten ergriffen sein Gewand
(T 203,1)
„Nachdem sie ihn erhängt hatten, ergriffen sie sein Gewand (…)"
vgl. Lat. Tatian:
Postquam autem crucifixerunt eum, acceperunt vestimenta eius …

Die Bestandteile derartiger Konstruktionen werden in den älteren Wörterbüchern als Einzelformen aufgefasst, so dass nur die Einträge für die einzelnen Bestandteile durch die automatisierte Vorannotation erstellbar sind.

Für *aftar thiu* in (1) werden nach der automatisierten Vorannotation, in diesem Fall mit SEHRT (1966), folgende Werte angeführt:

Wortform	*after*	*thiu*				
Lemma	aftar	the	the	the	thiu	thiwa
Übersetzung	zeitlich	der	der	der	Magd, Dienerin	Dienerin, Magd
PoS Lemma	AP	DD	DD	DD	NA	NA
PoS Beleg	APPR?	DDA	DDA	DDA	NA	NA
Flexion Lemma					jo_Fem	n_Fem
Flexion Beleg 1					jo_Fem	n_Fem
Flexion Beleg 2		Fem Sg Nom	Masc¦Neut Sg Ins	Neut Pl Nom¦Acc	Sg Nom,Acc	Sg Nom

Tabelle 2: Vorschläge für as.'after thiu' nach SEHRT (1966)

Anders erscheint die Vorannotation von *aftar thiu* im ahd. Tatian wie in (2) nach dem Glossar von SIEVERS (1892):

Wortform	*after*		*thiu*
Lemma	after	after	dër
Übersetzung	hinten, hinter … her, von hinten, nach hinten; (da)nach, später, entsprechend. m. Gen.: nach. m. Dat.: entlang, über … hin, hinter … her, auf, durch, nach, zu; hinsichtlich, entsprechend, zufolge,	hinten, hinter … her, von hinten, nach hinten; (da)nach, später, entsprechend. m. Gen.: nach. m. Dat.: entlang, über … hin, hinter … her, auf, durch, nach, zu; hinsichtlich, entsprechend, zufolge,	dieser, der(selbe); diese, die; dieses, das; wer, welcher; welche; was, welches

	gemäß. m. Akk.: hinter, nach, gemäß. subst.: das Hinten	gemäß. m. Akk.: hinter, nach, gemäß. subst.: das Hinten	
PoS Lemma	ADV	ADV¦AP	DD
PoS Beleg	ADV	ADV¦APPR?	DDA
Flexion Lemma			
Flexion Beleg 1			
Flexion Beleg 2			_Sg_Gen

Tabelle 3: Vorschläge für ahd.'after thiu' nach SIEVERS (1892)

Beide Beschreibungen entsprechen allerdings nicht unserem Verständnis derartiger Konstruktionen, wonach *aftar thiu* – und ähnliche Verbindungen – als eine Worteinheit aufgefasst und entsprechend annotiert werden.

Einen Eintrag aber über die möglichen Verbindungen von *thiu* in der Funktion eines Adverbs oder einer Subjunktion sucht man in den meisten Glossaren zu den altdeutschen Texten häufig vergeblich. Werden Verbindungen von einer Adposition mit *thiu* angeführt, erscheinen diese unvollständig und unsystematisch. Zudem wird meist lediglich eine Bedeutungsvariante genannt, jedoch wird dieser keine eigenständige grammatische Funktion zugewiesen.

Dabei tritt gerade *thiu* sehr häufig in verschiedenen Verbindungen auf, in denen eine demonstrative Funktion, wie die älteren Glossare *thiu* auch in diesen Fügungen wenigstens implizit geben, kaum noch zugewiesen werden kann und es sich oft auch formal nicht um einen Instrumentalis handeln kann.[6] In diesen Fällen werden Verbindungen aus *thiu* und Adposition vom Referenzkorpus Altdeutsch als lexikalisiert angesehen.

Für das vorliegende Korpus werden sämtliche Einzelwörter einer mehrteiligen Subjunktion oder eines Adverbs zusammengeführt und als eine Wortform annotiert.

4 Eine Frage der Perspektive: Lücken in der Beschreibung

Ein herausragendes Anliegen des Referenzkorpus Altdeutsch ist es, ein möglichst genaues Bild vom tatsächlichen Sprachstand der ältesten Überlieferung des Deutschen zu geben. Dieses Vorhaben erscheint trivial, erweist sich aber bei der zeitlichen und räumlichen Streuung der Texte und vor dem Hintergrund der grammatischen Beschreibungen der relevanten Glossare, auf die sich das Referenzkorpus Altdeutsch stützt, als relativ vielschichtige Aufgabe.

Die altdeutschen Sprachen zeigen eine relativ große Anzahl an substantivischen Deklinationsklassen auf, die jedoch zahlreichen Wandelerscheinungen unterliegen: Neben Kasusschwund (Instrumentalis) und formalem Kasuszusammenfall (z.B. Genitiv und Dativ Singular bei den o-Feminina) fallen verschiedene Deklinationsklassen zusammen oder gehen ineinander über, und es besteht bei vielen Substantiven eine Tendenz zum Genuswechsel.

Die Glossare greifen diese Heterogenität in Teilen auf, indem sie zwar von den ursprünglichen germanischen Deklinationsklassen ausgehen, jedoch abweichende Genera und Stammbildungen verzeichnen. Allerdings beschränken sich die Wörterbücher auf die Angaben von starker und schwacher Deklination, ohne die entsprechenden Stämme anzuführen. Hier ergänzen für die automatisierte Vorannotation die Grammatiken (BRAUNE 2004, GALLÉE 1993) die einzelnen Stammklassen. Diese zeigen grundsätzlich eine diachrone Anlage in der Einteilung der Deklinationsklassen ausgehend von der germanischen Stammbildung. So wird z.B. das Lemma *sunta* ‚Sünde' entsprechend dem ursprünglichen Thema als *jo*-Stamm angeführt.[7]

Die Belege zeigen aber, dass neben den j-haltigen Formen auch Flexionsformen auftreten, die den o-Stämmen entsprechen, dass also das Paradigma von *sunta* sich den o-Stämmen formal angleicht und damit die Voraussetzungen für einen Klassenwechsel (oder Klassenzusammenfall) gegeben sind.[8]

(7)	*dir*	*uuirdu*	*ih*	*pigihtik*	*allero*	*minero*	*suntiono*
	dir	werde	ich	„beichtig"	all-GEN.PL.	mein-GEN.PL.	Sünde-GEN.PL.
	(AB,1)						

(8) *dir uuirdu ih pigihtik allero mindero suntono*
(BG1,1)
„Ich beichte dir alle meine Sünden"

Im Vergleich dazu die Genitiv-Formen ursprünglicher o-Stämme:

(9) [Fortsetzung von (7)]

meinsuartio	*enti*	*lugino,*	*kiridono*	*enti*	*unrehteru*
Meineide-GEN.PL	und	Lügen-GEN.PL.	Begierde-GEN.PL.	und	unrechte-GEN.SG.
i-Stamm		a-Stamm	o-Stamm		

fizusheiti,	*huorono*
Hinterlist-GEN.SG.	Ehebruch-GEN.PL.
i-Stamm	o-Stamm
(AB,5)	

[ich beichte:]
„Meineide und Lügen, Begierden und unrechte Hinterlist, Ehebrüche"

Wenn wir also den Wandel bei den Deklinationen der Substantive abbilden wollen, ist es notwendig, die synchron im Altdeutschen tatsächlich auftretenden Formen zu kennzeichnen. Demnach werden Formen, die noch Reflexe des ursprünglichen Themas zeigen, als jo-Feminina gekennzeichnet, während abweichend von den Grammatiken alle anderen starken Formen auf der Belegebene als o-Feminina getaggt werden.

(10)	*dir uuirdu ih pigihtik allero minero*	*suntiono*	(AB,1)
	Lemma	sunta	
	Flexion L	jo_Fem	
	Flexion B	jo_Fem	
(11)	*dir uuirdu ih pigihtik allero minero*	*suntono*	(BGa,1)
	Lemma	sunta	
	Flexion L	jo_Fem	
	Flexion B	o_Fem	

5 Grammatisches Gemenge: Inadäquate Beschreibungen aufgrund der automatisierten Vorannotation

Trotz sorgfältiger Prüfung bleibt es nicht aus, dass bei der automatisierten Vorannotation Fehler entstehen. Diese zeigen sich vor allem bei Formen mit mehreren grammatischen Funktionen oder bei unterschiedlichen Lemmata, die formgleich sind.

Das ad. Lemma *filo* ‚viel, sehr' geht auf ein u-Neutrum *filu* ‚Vieles, Vielheit' zurück, welches in den altdeutschen Texten zum einen in Argumentposition als indeklinables Substantiv, zumeist mit abhängigem Genitiv (12), und zum anderen als graduelles Adverb (13) auftritt.

(12) *so sculun gi undar iuuua fiund faren, undar filu theodo*
so sollt ihr unter eure Feinde gehen unter 'Viel'-ACC.SG. Volk-GEN.PL.
(H 1875)
„So sollt ihr unter eure Feinde gehen, unter viele Völker (=unter eine Vielheit der Völker)"

(13) *thaz líoht ist filu war thing* (O II 2, 13)
das Licht ist viel wahr Ding
„Das Licht ist eine sehr wirkliche Sache"

In den ahd. Textwörterbüchern wird *filo* unterschiedlich beschrieben, in den meisten Fällen werden beide Varianten mit den entsprechenden Belegstellen genannt. Ausgehend von SPLETT (1993) für das Gesamtalthochdeutsche wurde es jedoch als Adverb vorannotiert, auch wenn dies nicht immer der tatsächlichen Lesart entspricht. In den Fällen, in denen es als Substantiv auftritt, muss deswegen die Vorannotation korrigiert werden.

Eine Analyse als Adjektiv wie in GALLÉE (1993:228) für das Altsächsische ist zumindest fraglich, jedoch in die Vorannotation aufgenommen. Für den Heliand annotiert das Referenzkorpus Altdeutsch in Übereinstimmung mit SEHRT (1966:180) *filo* als Adverb oder Substantiv, so dass auch hier von der Vorannotation abgewichen wird.

Eine andere Problematik birgt die Vorannotation der Homonyme ahd. *noh* a) ADV ‚noch, auch, außerdem' und b) KONJ ‚und nicht, auch nicht'. Aufgrund der Formengleichheit werden beide Lemmata in der Vorannotation zusammengefasst, in einigen Texten nur in den Fällen, bei denen die Lesart als Konjunktion möglich ist, in einigen Texten aber in allen Fällen. Dies ist von den Angaben im jeweiligen Textwörterbuch abhängig, das der Vorannotation zugrunde lag. Im Zweifelsfall muss der Annotator dann aufgrund des Kontextes entscheiden.

6 Fazit

Die Erfahrungen der Arbeiten des Referenzkorpus Altdeutsch zeigen, dass die zum Teil schon sehr alten Glossare zu den ältesten Texten des Deutschen und die Grammatiken für ein digital aufbereitetes, tiefenannotiertes Korpus von großem Nutzen sind. Die zunächst zeitaufwändige Methode, sämtliche Textwörterbücher jeweils zu den einzelnen Überlieferungen einzulesen und dann die daraus gewonnenen Informationen manuell im Annotationstool zu bearbeiten, erweist sich als ausgesprochen wirkungsvoll. Da so gut wie jede Wortform der altdeutschen Texte in den entsprechenden Glossaren detailliert verzeichnet ist, ergeben sich bei der lemmabezogenen Vorannotation nur dann Abweichungen, wenn die Bearbeiter/innen des Referenzkorpus Altdeutsch grundsätzlich eine Form anders als vorgeschlagen deuten. Diese Fälle können wie gezeigt systematisch behandelt und begründet werden.

---

[1] Das Forschungsprojekt Referenz Korpus Altdeutsch http://www.deutschdiachrondigital.de wird von der Deutschen Forschungsgemeinschaft (DFG) gefördert und ist am Lehrstuhl für Sprachgeschichte an der Humboldt-Universität zu Berlin, am Institut für Empirische Sprachwissenschaft an der Goethe-Universität Frankfurt a.M. und am Lehrstuhl für Indogermanistik an der Schiller-Universität Jena angesiedelt.

[2] http://www.lat-mpi.eu/tools/elan/

[3] HEFFNER (1961), HENCH (1890), HENCH (1893), KELLE (1881), SEHRT (1955), SEHRT (1966), SIEVERS (1874), SIEVERS (1892) und WADSTEIN (1899). Bisher standen keine digitalisierten Versionen zur Verfügung.

[4] Zur Kennzeichnung der PoS-Kategorien verwendet das Referenzkorpus Altdeutsch ein an das STTS (Stuttgart Tübingen Tagset, SCHILLER et al., 1999) angelehntes Tagset, welches dieses für die Annotation historischer Daten modifiziert. Ein Ausschnitt des DDDTS (Deutsch Diachron Digital Tagset) mit den in diesem Aufsatz aufgeführten Tags befindet sich im Anhang.

[5] BRAUNE/REIFFENSTEIN (2004), GALLÉE (1993).

[6] Dies erscheint allenfalls noch in Verbindung mit der Präposition as. *mid*, ahd. *mit* , die u.a. den Instrumentalis fordert, möglich; in der Bedeutung eines Pronominaladverbs *damit* . Vgl.: *endi fôdie is hundos mid thiu* ‚und füttere die Hunde damit' (Hel 3017), wobei *thiu* sich hier auf das vorerwähnte Neutrum *brôd* ‚Brot' bezieht und damit formal einem substituierenden Demonstrativpronomen entspricht.

[7] Das germ. Substantiv besteht ursprünglich aus drei Morphemen: Wurzel, Stammbildungselement (Thema) und Flexionsendung. Die Wurzel trägt die inhaltliche Bedeutung und ist zunächst eine grammatisch neutrale Abstraktion, welche normalerweise nicht isoliert auftritt. Aus der Wurzel werden durch Stammbildungselemente Lexeme abgeleitet, welche in der Regel der Wurzel eine grammatische Kategorie zuweisen. Wurzel und Stammbildungselement bilden den Wortstamm, welcher nicht alleine auftreten kann, sondern eine Flexionsendung erhalten muss. Die Einteilung in Stammklassen erfolgt für das Ahd. aufgrund des ursprünglichen Themas, auch wenn dieses häufig nicht mehr sichtbar ist. Allerdings zeigen sich in vielen Formen noch die Reflexe der ursprünglichen Stammbildungselemente, die dann üblicherweise als Teile der Endung interpretiert werden. So zeigt Nom.Sg. *suntea* ‚Sünde' mit dem *-e-* noch einen Rest des Themabestandteils *-j-,* während in der Form Nom.Sg. *sunta* dieses geschwunden (oder vollkommen verschmolzen) ist.

[8] Nach dieser Auffassung werden die ursprünglichen Stammbildungssuffixe synchron Althochdeutsch und Altniederdeutsch als Teil des Flexionsmorphems betrachtet; die Umlaute werden spätestens für den Übergang zum Mittelhochdeutschen als nicht mehr phonologisch determiniert angesehen. Die Annahme der Reanalyse des Themas als Flexiv und der Lexikalisierung des Umlauts bildet eine der Voraussetzungen für den möglichen Zusammenfall von substantivischen Deklinationsklassen im Deutschen. Beim Tagging der altdeutschen Daten wird von dieser Annahme ausgegangen, um morphologische Schwankungen präzise darstellen zu können.

Anhang: DDDTS (Deutsch Diachron Digital Tagset)
Ausschnitt

PoS Beleg	PoS Lemma	Beschreibung
DPOS	DPOS	Determinativ, possessiv
DPOS	PPER	Sonderfall: Genitiv des PPER als DPOS
NA	NA	Nomen, Appellativ
PPER	PPER	Pronomen, personal, irreflexiv
VVFIN	VV	finites Verb, Vollverb
VVINF	VV	Infinitiv, Vollverb
VVPS	VV	Partizip Präsens, Vollverb, im Verbalkomplex
VVPSA	VV	Partizip Präsens, Vollverb, attribuierend

Literatur

Textausgaben

AB Altbairische Beichte. In: E. Steinmeyer (Hg.). (1971). Die kleineren althochdeutschen Denkmäler. Dublin, Zürich: Weidmann, 309. (Nachdruck von 1916).

BGa Bairisches Gebet A. In: E. Steinmeyer (Hg.). (1971). Die kleineren althochdeutschen Denkmäler. Dublin, Zürich: Weidmann, 310-314. (Nachdruck von 1916).

H Heliand. (1984). Hg. Burkhard Taeger, Tübingen: Niemeyer. 9. Aufl.

O Otfrids Evangelienbuch. (1973). Hg. Oskar Erdmann, Tübingen: Niemeyer. 6. Aufl.

T Tatian. (1966). Hg. Eduard Sievers. Paderborn: Schöningh. (Nachdruck der zweiten Ausgabe von 1892).

Grammatiken und Wörterbücher

Braune, W./Reiffenstein, I. (2004). Althochdeutsche Grammatik. Tübingen: Niemeyer. 15. Aufl.

Gallée, J. H. (1993). Altsächsische Grammatik. Tübingen: Niemeyer. 3. Aufl.

Heffner, R.-M. S. (1961). A Word-Index to the Texts of Steinmeyer. Die kleineren althochdeutschen Sprachdenkmäler. Madison: The University of Wisconsin Press.

Hench, G. A. (1890). The Monsee Fragments. Strassburg: Trübner.

Hench, G. A. (1893). Der althochdeutsche Isidor. Strassburg: Trübner.

Kelle, J. (1881). Glossar der Sprache Otfrids. Regensburg: Manz.

Sehrt, E. (1955). Notker-Wortschatz. Halle: Niemeyer.

Sehrt, E. (1966). Vollständiges Wörterbuch zum Heliand und zur altsächsischen Genesis. Göttingen: Vandenhoeck & Ruprecht.

Sievers, E. (1874). Die Murbacher Hymnen. Halle: Buchhandlung des Waisenhauses.

Sievers, E. (1892). Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar. 2nd edition. Paderborn: Schöningh.

Splett, J. (1993). Althochdeutsches Wörterbuch. Berlin: de Gruyter.

Wadstein, E. (1899). Kleinere altsächsische Sprachdenkmäler. Norden/Leipzig: Soltau.

Dennis Pauly, Ulyana Senyuk, Ulrike Demske

Strukturelle Mehrdeutigkeit in frühneuhochdeutschen Texten

1 Einleitung

In der synchron orientierten Sprachwissenschaft ist in letzter Zeit ein verstärktes Interesse an syntaktisch annotierten Korpora zu erkennen. Für das Gegenwartsdeutsche wären hier das TIGER- bzw. NEGRA-Projekt zu nennen, in deren Rahmen bereits umfangreiche, syntaktisch annotierte Zeitungskorpora entstanden sind (BRANTS et al. 1999, 2002). Gerade in früheren Sprachstufen sind digitalisierte und linguistisch aufbereitete Korpora als Datenquelle allerdings von noch größerer Relevanz, denn eine introspektive Datenerhebung ist hier im Gegensatz zum Gegenwartsdeutschen nicht möglich, eine Untersuchung hat also notwendigerweise immer korpusbasiert zu erfolgen. Dementsprechend gibt es für bestimmte Sprachstufen wie das Mittelenglische mit der Penn-Datenbank schon größere, syntaktisch annotierte Korpora (KROCH/TAYLOR 2000). Auch für die einzelnen Sprachperioden des Deutschen werden derzeit Referenzkorpora aufgebaut, die mit linguistischen Informationen angereichert sind. Dazu gehören neben den DFG-Projekten zum Althochdeutschen (Berlin, Frankfurt/M. und Jena), Mittelhochdeutschen (Bochum und Bonn) und Frühneuhochdeutschen (Bochum, Halle und Potsdam) auch das GerManC-Projekt zum frühen Neuhochdeutschen (1650 – 1800) an der Universität Manchester und das ISWOC-Projekt (Information Structure and Word Order Change in Germanic and Romance Languages, BECH/EIDE 2011), das unter anderem syntaktische Informationen zum Althochdeutschen enthalten wird. Das hier vorgestellte Korpus ist ein syntaktisch annotiertes Korpus des Frühneuhochdeutschen, das im Rahmen eines Pilotprojekts von 2003 bis 2005 an der Universität des Saarlandes mit dem Ziel entstanden ist, an Texten, die sich sowohl durch große Varianz auf allen Ebenen des Sprachsystems als auch durch eine große Komplexität ihrer Phrasen und Sätze auszeichnen (ADMONI 1980), die Möglichkeiten einer halbautomatischen Annotation zu erproben. Basierend auf den Erfahrungen aus diesem Pilotprojekt sollen dann größere Textmengen aus dem Frühneuhochdeutschen im Baumbankformat aufbereitet und als annotiertes Referenzkorpus auf einer geeigneten Plattform frei zugänglich zur Verfügung gestellt werden. Eine solche Baumbank historischer Texte ermöglicht es dann, ausgesuchte Fragestellungen der historischen Syntax gezielter und auch in quantitativer Hinsicht zu untersuchen.[1] Darüber hinaus stellt die hohe Komplexität aus annotatorischer Sicht auch eine besondere Herausforderung dar, was die Qualität bzw. Konsistenz der Annotation angeht.

Wir werden im Folgenden das syntaktisch annotierte MERCURIUS-Korpus zum Frühneuhochdeutschen vorstellen und dabei sowohl auf die Textauswahl wie auch auf die gewählte Annotationsweise näher eingehen. Anhand von morphologischen Strukturen wie N-N-Komposita und Partikelverben sollen dann exemplarisch die Probleme disku-

tiert werden, die sich durch die Herausbildung dieser Lexeme aus syntaktischen Phrasen für die Annotation in den frühneuhochdeutschen Texten ergeben, und ein möglicher Umgang mit Problemen dieser Art aufgezeigt werden.

2 Eine Baumbank frühneuhochdeutscher Zeitungstexte

2.1 Korpus

Das MERCURIUS-Korpus (DEMSKE 2007) – benannt nach einem der enthaltenen Texte – besteht aus bislang zwei Jahrgängen frühneuhochdeutscher Zeitungstexte, welche mit syntaktischen Informationen angereichert sind. Dabei handelt es sich um den 'Nordischen Mercurius' (M) von 1667 und die Monatsschrift 'Annus Christi' (AC), die 1597 erschienen ist. Damit stehen zwei Texte zur Verfügung, die beide aus einer eher späten Phase des Frühneuhochdeutschen stammen. Trotz der immer noch großen Varianz, die für diese Periode der deutschen Sprachgeschichte charakteristisch ist, erleichtert die Nähe zum Gegenwartsdeutschen den Einsatz eines Annotationsschemas, das für das Gegenwartsdeutsche entwickelt wurde. Das annotierte Korpus umfasst bislang etwa 170.000 Wortformen, wobei knapp über 130.000 Wortformen bzw. 7.500 Sätze aus dem Nordischen Mercurius und etwas mehr als 40.000 Wortformen und 1.000 Sätze aus dem Annus Christi stammen. Damit liegt für das Frühneuhochdeutsche ein erstes syntaktisch annotiertes Korpus vor, mit dem nicht nur qualitative, sondern auch quantitative Auswertungen mit vertretbarem Aufwand möglich sind.

2.2 Annotation

Bevor eine Annotation der jeweiligen Texte geleistet werden konnte, mussten die vorhandenen Zeitungsjahrgänge digitalisiert und segmentiert werden.[2] Die syntaktische Annotation erfolgte mittels des Werkzeugs 'Annotate', welches zur Annotation gegenwartsdeutscher Texte im Rahmen der TIGER- bzw. NEGRA-Baumbank-Projekte entwickelt wurde (BRANTS et al. 1999, 2002). Das hybride Annotationsschema TIGER, das hierbei zur Anwendung kommt, unterscheidet zwischen drei syntaktischen Annotationsebenen, nämlich der Wort-, Phrasen- und Funktionsebene. Die Annotation der Wortarten erfolgt nach dem STTS (Stuttgart-Tübingen-Tagset, vgl. SMITH/EISENBERG 2000). Dem folgt – wie in Abb. 1 zu sehen ist – die Auszeichnung der Phrasenstruktur, welche rund umrahmt dargestellt wird, und schließlich die Einbindung einer funktionalen Ebene (eckig umrahmt). Die Annotation selbst wird halbautomatisch vorgenommen, was in diesem Fall heißt, dass dem Annotator die Lösungsvorschläge unterbreitet werden, die mit der höchsten Wahrscheinlichkeit zutreffen. Diese Vorschläge beruhen auf dem Markov-Modell (BRANTS 1999), welches mit statistischen Daten arbeitet, die aus bereits annotierten Teilen des Korpus stammen. Auf diese Weise werden zum einen auf der Wortartebene und zum anderen auf der phrasalen Ebene Vorschläge für potentiell richtige Strukturen unterbreitet. Bei den Vorschlägen auf der Wortartebene kann der Annotator entweder den unterbreiteten Vorschlag annehmen oder einen anderen Vorschlag aus einer nach Wahrscheinlichkeit geordneten Liste auswählen. Was die Vorschläge auf der

phrasalen Ebene angeht, kann der Annotator die vorgeschlagene Phrasenstruktur entweder akzeptieren oder sie entsprechend modifizieren. Während der Annotation wird dem Annotator jeweils ein vollständiger Satz präsentiert. Er kann jedoch jederzeit im Korpus blättern und vorausgehende oder nachfolgende Sätze ansehen. Die Annotation selbst wird stets von zwei Personen vorgenommen, die im sog. 'double keying'-Verfahren jeweils denselben Teiltext selbständig annotieren und danach ihre Versionen computergestützt miteinander abgleichen. Auf diese Weise kann mit einem Mehr an Arbeitsaufwand höchstmögliche Konsistenz gewährleistet werden, wie BRANTS (2000) gezeigt hat.

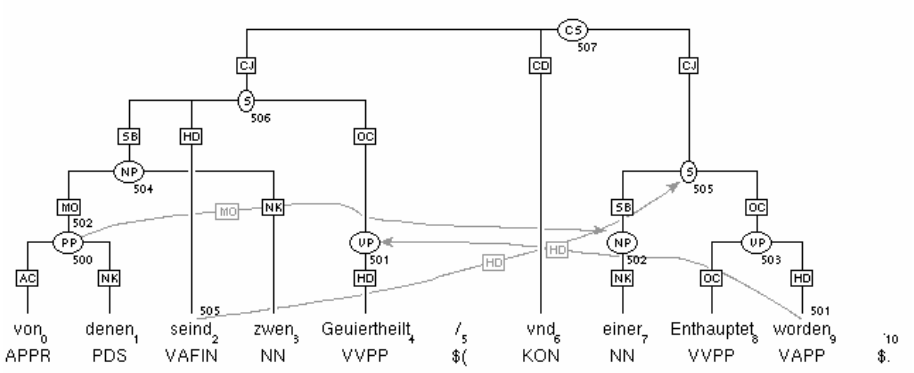


Abb. 1 Annotationsbeispiel

Ein wichtiges Merkmal ist weiterhin, dass versucht wurde, die Baumdiagramme möglichst übersichtlich zu gestalten und trotzdem alle wichtige Informationen zu integrieren. Diese Herangehensweise ist an den flachen Baumstrukturen zu erkennen. Ein Beispiel für eine solche Struktur stellt die Präpositionalphrase *von denen* in Abb. 1 dar. Präposition und nominaler Kopf sind Schwesterknoten einer Phrase, so dass Phrasen vom Typ PP problemlos gefunden werden können. Informationen über die interne Struktur der Präpositionalphrase enthält die flache Annotation dagegen nicht. Diese Vereinfachung erlaubt jedoch ein Annotieren größerer Textmengen, da die Bearbeitungszeit durch die Annotatoren wesentlich verkürzt werden kann.

Ferner besteht die Möglichkeit, sich überkreuzende Kanten wie auch sog. sekundäre Kanten zu erstellen, welche es ermöglichen, diskontinuierliche oder elliptische Strukturen darzustellen, die im Frühneuhochdeutschen häufig zu finden sind. So wird in der Abb.1 mit der sekundären Kante, die vom Verb *seind* zum Satzknoten S führt, angezeigt, dass dieses Verb semantisch und strukturell nicht nur in den ersten, sondern – wenn auch nicht in dieser Flexionsform[3] – ebenso in den zweiten Satz gehört, obwohl es an dieser Stelle ausgelassen wurde. Entsprechende Informationen werden auch durch die sekundären Kanten vom Verb *worden* und von der Präpositionalphrase *von denen* erfasst. Mit dem innerhalb des TIGER-Projekts geschaffenen Suchwerkzeug TIGERSearch kann

schließlich sichergestellt werden, dass speziell auf syntaktische Phänomene ausgerichtete Suchanfragen bearbeitet werden können (LEZIUS 2002, VOORMANN/LEZIUS 2002).[4]

3 Strukturelle Mehrdeutigkeit

In diesem Abschnitt soll es um ein Problem gehen, das sich verstärkt bei der Annotation historischer Texte stellt und den Zusammenhang von synchroner Variation und Sprachwandel betrifft. Im Besonderen geht es um die Frage, wie bei der Anreicherung eines Textes mit syntaktischen Informationen mit sprachlichen Mustern umzugehen ist, für die aufgrund des Kontextes nicht entschieden werden kann, ob es sich um die Ausgangs- oder die Zielstruktur eines sprachlichen Wandels handelt. Diese Frage soll im Folgenden exemplarisch für Univerbierungsprozesse im Deutschen diskutiert werden, die der Entstehung von N-N-Komposita und Partikelverben zugrunde liegen.

3.1 Uneigentliche Komposita

Wortbildungsmuster mit vermeintlichen Genitivformen als Erstglied sind im Gegenwartsdeutschen sehr verbreitet (1). Dass es sich bei der linken Komponente jedoch nicht um eine Genitivform handelt, zeigen Daten wie unter (1b): Feminine Nomina, die kein *-(e)s* in ihrem Paradigma haben, unterstützen die Annahme, dass es sich lediglich um eine morphologisch gesehen neutrale Fuge handelt und nicht um ein Flexionssuffix (RAMERS 1997).

(1) a. Gottesbeweis
b. Liebesbeweis

Demgegenüber stehen ähnlich aussehende, jedoch phrasal zu interpretierende Formen mit pränominalem Genitiv (2). Genitivische Nominalphrasen können ihrem Kopfnomen vorangehen, wenn es sich um Eigennamen handelt (2a), bestimmte Pronomina (2b) oder – wie in (2c) – um belebte Individualnomina, wobei letztere Variante markiert ist. Sicher noch stärker markiert ist die pränominale Position für unbelebte Individualnomina (2d).

(2) a. Julias Telefon
b. dessen Buch
c. ?des Königs Kleider
d. ??des Briefes Herkunft

Diese einfache Unterscheidung zwischen morphologischer und syntaktischer Struktur ist im Frühneuhochdeutschen jedoch häufig nicht möglich. So kann beispielsweise für Ausdrücke wie (3) nicht entschieden werden, ob es sich um eine syntaktische oder eine morphologische Struktur handelt. Hier könnte entweder eine durch einen pränominalen Genitiv modifizierte Nominalphrase (3a) oder ein Nominalkompositum (3b) vorliegen.

(3) Kriegß Expedition[5] (AC 177)[6]

a. [NP [NP Kriegß] Expedition]
b. [N [N Kriegß] [N Expedition]]

Die Getrennt- bzw. Zusammenschreibung bestimmter Strukturen ist dabei nur bedingt aussagekräftig. Oft liegen nämlich identische Strukturen in zweierlei Schreibung vor (4). Dementsprechend wird auch davon ausgegangen, dass sich die oben erwähnten Komposita per Reanalyse aus einem entsprechenden syntaktischen Muster entwickelt haben (Demske 2001, 305).

(4) a. KriegsWaffen (M 7153)
b. Kriegs Waffen (M 209)

Erschwerend kommt in diesem Fall hinzu, dass die für das Gegenwartsdeutsche erwähnten Beschränkungen hinsichtlich des Vorkommens pränominaler Genitive für das Frühneuhochdeutsche nicht gelten. Das heißt, dass es im Frühneuhochdeutschen grundsätzlich möglich ist, Nomina aller semantischer Subklassen als pränominale Modifikatoren zu benutzen. Theoretisch ist deshalb sowohl eine Analyse des Belegs unter (3) als morphologische wie auch als syntaktische Struktur möglich.

Bevor wir jedoch unsere Herangehensweise an dieses Annotationsproblem vorstellen, soll zunächst eine Übersicht über die im Korpus belegten Typen von Kompositionsbildungen gegeben werden. Aus (5) wird ersichtlich, dass nicht nur Nomina als 'Erstglieder' infrage kommen, sondern auch Vertreter anderer Wortarten. So ist das determinierende Element in (5a) ein Adverb, in (5b) eine Präposition und in (5c) ein Verb. Auch Negationspartikeln (5d), Konfixe (5e) bzw. Adjektive (5f) sind als Konstituenten von Komposita vertreten.

(5) a. wider eroberung (AC 339)
b. Vor Jahre (M 435)
c. Aufziech Brucken (AC 414)
d. nicht annehmung (M 1596)
e. Vice Cancellario (M 6198)
f. Groß Vezier (M 2205)

Darüber hinaus sind die möglichen Kompositionsbildungen nicht auf Nominalkomposita beschränkt. Genauso finden sich nämlich auch Adverbien, welche durch Partikeln (6a) erweitert sein können, und auch in Verbindung mit Adjektiven kommen entsprechende Muster vor, in denen Zahlwörter (6b), Adjektive (6c), Verben (6d) oder Nomina (6e) als Modifikatoren links vom adjektivischen Kopf erscheinen.

(6) a. als baldt (AC 224)
b. sechs tägliche Robat (AC 33)
c. nechst gelegne Spital (AC 44)
d. glaub würdig (M 1042)
e. Mast und Segel loß (AC 118)

Was die Annotation angeht, so sind die meisten der oben erwähnten Daten unproblematisch. In nahezu allen Fällen mit einem Adjektiv oder Adverb als Kopf muss von einer Kompositionsbildung ausgegangen werden, da sich eine syntaktische Modifikationsbeziehung schnell ausschließen lässt. So ist es bspw. ausgeschlossen, dass Adjektive durch Verben modifiziert werden (6d).

Allerdings gibt es auch schwieriger einzuordnende Fälle: Als Erstes möchten wir nochmals auf die anfangs genannte Struktur (3) – hier wiederholt als (7) – zurückkommen.

(7) Doch begehren die Stände / daß zu dieser **Kriegß Expedition**, so wol auf die Gränitzhäuser zu Obristen / Rittmeistern / Hauptleuten vnnd andern Befelchshabern / geboren Behaimb / oder auß derer zu diesem Königreich gehörende / Länder / bürtige Personen dazu daugentlich / vor andern gefürdert vnnd gebraucht werden möchten (AC 177)

a. [$_{NP}$ [$_{NP}$ Kriegß] Expedition]
b. [$_{N}$ [$_{N}$ Kriegß] [$_{N}$ Expedition]]

Wir nehmen an, dass eine komplexe Phrase ausschließlich dann vorliegt, wenn mindestens eine der beiden folgenden Bedingungen zutrifft:

(i) Das Erstglied wird in dem betreffenden Fall modifiziert respektive spezifiziert.

(ii) Es liegt keine Adjazenz zwischen Erstglied und potentiellem Kopf vor.

Infolgedessen muss für Fälle wie (7) angenommen werden, dass es sich um ein komplexes Wort und keine komplexe Phrase handelt. Genau entgegengesetzt müssen hingegen Daten wie (8) gesehen werden. Hier können nur phrasale Strukturen vorliegen: In (8a) wird das Nomen *Türcken* durch einen Determinierer spezifiziert, während in (8b) und (8c) sogar beide Bedingungen für das Vorliegen einer komplexen Phrase erfüllt sind. Zum einen wird das jeweils linke Nomen durch einen Determinierer spezifiziert, zum anderen ist keine Adjazenz gegeben, denn in (8b) interveniert die Präpositionalphrase *auß spania* zwischen den beiden Nomina, in (8c) der adjektivische Modifikator *grosser*.

(8) a. die darinn gelegne Heiducken aber [...] **der Türcken Obristen** sampt noch 13 zu Boden gelegt [...] (AC 567)
b. in des Königs auß spania Gewalt (M 537)
c. der Türcken grosser Verlust (M 5575)

In allen anderen Fällen, die nicht der Form ‘N + N’ entsprechen, wird entsprechend verfahren. So könnte in (9a) grundsätzlich auch ein attributives Adjektiv und damit eine komplexe Phrase vorliegen. Das fehlende Flexionssuffix kann hierbei nicht als hinreichendes Kriterium für eine Kompositionsanalyse betrachtet werden, denn im Frühneuhochdeutschen fehlen overte Kasusmarkierungen häufig. Ebenso könnte man für (9b) annehmen, dass das linke Adjektiv zusammen mit dem rechten eine komplexe Phrase bildet, zumal im Gegenwartsdeutschen bei ähnlich gelagerten Ausdrücken beide Interpretationen möglich sind, was sich in der Schreibung widerspiegelt (siehe z.B. *nahe*

gelegen vs. *nahegelegen*). Solche Fälle werden in der Baumbank jedoch analog zu den oben formulierten Bedingungen (i) und (ii) als Kompositionsbildungen behandelt.

(9) a. Groß Vezier (M 2205)
b. nechst gelegne Spital (AC 44)

Ein ebenfalls mehrdeutiger Fall, der Ähnlichkeiten mit (7) aufweist, aber noch schwieriger zu beurteilen ist, liegt in (10) vor. Dadurch, dass auch das Kopfnomen *Tages* im Genitiv steht, kann aufgrund der Oberflächenstruktur nicht entschieden werden, ob sich der Determinierer auf das rechte oder auf das linke Nomen in der Nominalphrase bezieht. Es ist also anhand des Beispiels nicht zu klären, ob Bedingung (i) vorliegt oder nicht. In Abhängigkeit von der gewählten Interpretation kann die Analyse aussehen wie in (10a) bzw. (10b).

(10) des Reichs Tages (M 105)
a. [$_{\text{NP}}$ [$_{\text{NP}}$ des Reichs] Tages]
b. [$_{\text{NP}}$ des [$_{\text{N}}$ Reichs Tages]]

Für diese Fälle gehen wir in der MERCURIUS-Baumbank davon aus, dass sich der Determinierer auf das Kopfnomen bezieht, der Sprachwandel also vollzogen ist. Dies bringt sowohl positive als auch negative Aspekte mit sich: Zum einen kann so zwar nicht garantiert werden, dass die vorgeschlagene Lösung für wirklich jedes Datum philologisch angemessen ist. Dies ist jedoch auch nicht unser Anspruch innerhalb der Korpusaufbereitung: Die Baumbank soll vielmehr als Datenquelle für die philologische Analyse bestimmter (ambiger) Strukturen dienen. Deshalb ist es wichtig, dass den Annotatoren klare Kriterien hinsichtlich der Annotation der Daten zur Verfügung stehen und dass alle einschlägigen Fälle mit einer TIGERSearch-Suche gefunden werden können. Die Auffindbarkeit kann in diesem Fall mittels der Suche nach dem dazugehörigen Label, das gleich vorgestellt werden soll, gewährleistet werden. Zum anderen wird so – und dies ist, was die Annotation angeht, der maßgebliche Punkt – eine konsistente und zügige Annotation gewährleistet.

Kommen wir nun zur technischen Seite der beiden unterschiedlichen Analysen: Eine Annotation als pränominaler Genitiv kann innerhalb des TIGER-Schemas problemlos geleistet werden. Dazu wird die jeweilige Phrase mit der Funktion GL (= Genitiv links) in die jeweilige Nominalphrase eingehängt (siehe hierzu Abb. 2). Liegt dagegen trotz Spatium eine morphologische Struktur vor, so liefert das Annotationsschema kein eigenes Label für die Auszeichnung der linken Wortteilkonstituente. Wir haben uns deshalb zur Einführung eines neuen Labels entschlossen. Mittels KOMPE (= Kompositions-Erstglied) kann auf Wortartebene markiert werden, dass es sich beim Kern der betroffenen Phrase um eine Kompositionsbildung handelt (vgl. Abb. 3). Damit sind die einschlägigen Belege im Korpus für entsprechende Recherchen identifizierbar.

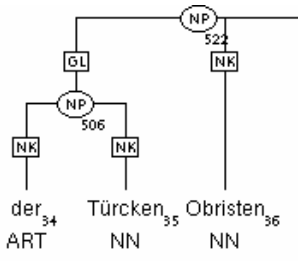


Abb. 2 Phrasale Struktur

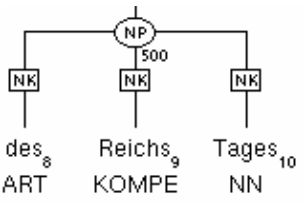


Abb. 3 Komposition

3.2 Partikelverben

Etwas anders gelagert sind die Probleme, die sich bei der Annotation von Partikelverben ergeben, weshalb diese hier gesondert behandelt werden. Partikelverben sind morphologisch komplexe Verben wie *anrufen*, deren Erstglieder – die sog. Verbpartikeln – formal gesehen identisch mit Präpositionen (11a), Adverbien (11b), Adjektiven (11c) oder Nomina (11d) sind. Meistens kann jedoch nicht von der Bedeutung der selbständigen Wörter auf die Bedeutung der Verbpartikeln geschlossen werden (Duden 2009, Helbig/Buscha 2007).

(11) a. aufstehen ('sich erheben' oder 'wach werden')
b. zusammenschlagen ('jmdn. niederschlagen')
c. schönreden ('beschönigen')
d. teilnehmen ('bei etw. dabei sein')

Dadurch, dass Verbpartikeln formal wie frei vorkommende Wörter aussehen, kann es vor allem bei Partikelverben mit Adjektiven, Adverbien und Nomina als Erstglied zu einer Verwechslung mit korrespondierenden syntaktischen Strukturen kommen. Bei Präpositionen stellt sich dieses Problem nicht, da sie – wenn sie frei vorkommen – stets mit einem Komplement auftreten. Handelt es sich hingegen um deadjektivische Verbpartikeln wie in (12), müssen Bedeutungsunterschiede herangezogen werden, um zwischen morphologischer und syntaktischer Struktur zu differenzieren: In (12a) enthält das Partikelverb eine lexikalisierte bzw. idiomatisierte Bedeutung, die sich nicht kompositional aus den Partikelverbkonstituenten ergibt, vgl. *kaltstellen* und die Bedeutung 'verdrängen', 'beiseite schieben'. Bei der Adjektiv-Verb-Konstruktion in (12b) ist die Bedeutung transparenter, d.h. man kann aus den einzelnen Teilen *kalt* und *stellen* auf die Gesamtbedeutung dieser Konstruktion schließen.

(12) a. Er **stellte** seinen Konkurrenten **kalt**. (Verb + Verbpartikel)
b. Er **stellte** die Torte **kalt**. (Verb + prädikatives Adjektiv)

Ein weiteres Unterscheidungskriterium stellt im Gegenwartsdeutschen die Betonung dar. Dieses Kriterium wird meistens zur Unterscheidung von Verbpartikel und selbständigem Adverb gebraucht: Bei Partikelverben liegt der Akzent normalerweise auf der Verbpartikel (13a), während bei syntaktischen Strukturen entweder nur das Verb oder beide Teile – Adverb und Verb – betont sein können (13b).[7]

(13) a. Ich konnte alle meine Freunde **WIEdersehen**.
b. Er konnte nach der Augenoperation **wieder SEhen/ WIEder SEhen**.

Die mithilfe der oben beschriebenen linguistischen Kriterien herausgearbeitete Analyse spiegelt sich im Gegenwartsdeutschen in der Orthografie wider: Partikelverben werden zusammengeschrieben (14a), während Getrenntschreibung ein visuelles Zeichen dafür ist, dass es sich um syntaktische Strukturen handelt (14b).[8]

(14) a. Er hat seinen Konkurrenten **kaltgestellt**. (Partikelverb)
b. Er wollte die Torte **kalt stellen**. (Verb + prädikatives Adjektiv)

Im Unterschied zum Gegenwartsdeutschen bereitet die Unterscheidung zwischen Partikelverb und syntaktischer Struktur in frühneuhochdeutschen Texten ein Problem, weil die Orthografie in diesem Fall wie auch schon bei den oben besprochenen N-N-Komposita häufig keinen Hinweis darauf gibt, ob eine morphologische oder syntaktische Struktur vorliegt. Wie (15) zeigt, kann beispielsweise die Verbindung von Präposition und Verb sowohl zusammen- als auch getrenntgeschrieben vorkommen, obwohl es sich in diesem Fall wegen des fehlenden präpositionalen Komplements um ein Kompositum handeln muss.

(15) a. [...] der Bremischen Friedenshandlung **beyzuwohnen** (M 2174)
b. [...] den FriedensTractaten **bey zu wohnen** (M 1323)

Für die Fälle, in denen sich Verben mit Nomina, Adjektiven und Adverbien verbinden, stellt sich jedoch die Frage, wie bei fehlender sprachlicher Kompetenz entschieden werden kann, ob ein Partikelverb vorliegt oder nicht, da die Prosodie in den historischen Textkorpora als diagnostisches Kriterium ausscheidet und semantische Unterschiede oftmals schwierig zu erkennen sind. Dass es sich hier nicht um ein zu vernachlässigendes Problem weniger Einzelfälle handelt, ergibt sich aus der Beobachtung, dass viele im Gegenwartsdeutschen gebräuchliche Partikelverben erst im Verlauf der jüngeren deutschen Sprachgeschichte als Produkte von Inkorporation, also einem Prozess entstanden sind, bei dem ein Verbstamm einen anderen mit ihm in Beziehung stehenden Wortstamm morphologisch eingliedert, wodurch ein komplexes Verb gebildet wird (Eisenberg 2006, 234). So hat sich das komplexe Verb *standhalten* wohl ursprünglich aus der syntaktischen Objekt-Verb-Konstruktion *Stand halten* entwickelt. Viele komplexe Verben mit Adjektiven und Adverbien als Erstglieder wie *krankliegen, bereitstellen, hinaufgehen* sind ebenfalls aus einst syntaktischen Konstruktionen entstanden (Eisenberg 2006, 332ff). Für zahlreiche frühneuhochdeutsche Belege von Partikelverbkandidaten ist folglich zu entscheiden, ob der Inkorporationsprozess bereits abgeschlossen ist oder im Einzelfall

noch syntaktische Strukturen vorliegen. In Fällen wie (16) muss folglich geklärt werden, wie eine morphosyntaktische Annotation aussehen muss, die den sprachlichen Fakten gerecht wird, ohne deren Analyse vorwegzunehmen. Anders ausgedrückt muss es für jeden Nutzer der Baumbank möglich sein, in dem annotierten Korpus qua TIGERSearch alle für die Herausbildung von Partikelverben einschlägigen Varianten qualitativ und quantitativ zu erfassen. Gleichzeitig muss gewährleistet sein, dass den Annotatoren unzweifelhafte Kriterien für die Annotation von Mustern wie (16) an die Hand gegeben werden.

(16) [...] vor den Holländern wol **weg gekommen** sey (M 4248)

3.2.1 Annotation von Partikelverben: Generelles

Was die tatsächliche Vorgehensweise beim Annotieren angeht, so wird hier ähnlich wie bei den im vorausgehenden Abschnitt diskutierten Kompositionsfällen verfahren, mit der Ausnahme, dass eine zusätzliche Annahme getroffen werden muss. Von einer Verbpartikel und somit einer morphologischen Struktur sprechen wir also, wenn

(i) die mutmaßliche Verbpartikel und das Verb in der Grundstruktur adjazent stehen,[9]

(ii) keine Modifizierung bzw. Spezifizierung der mutmaßlichen Verbpartikel vorliegt (bei Nomen, Adjektiven oder Adverbien als Verbpartikel) bzw. bei der mutmaßlichen Verbpartikel kein Komplement auftritt (bei Präpositionen als Verbpartikel),

(iii) das fragliche Lexem als Partikelverb im heutigen Deutsch belegt ist.

Die Fälle in (17) werden somit als Partikelverben analysiert und das Erstglied mit dem Wortart-Tag PTKVZ (= abgetrennter Verbzusatz) versehen (vgl. dazu Abb. 4).

(17) a. [...] bald soll **mit getheilet** werden. (M 2794)
b. [...] darauff sie folgenden Tag **fort ziehen** wöllen / (AC 787)

Bei den Daten in (18) wird entsprechend der genannten Kriterien jedoch von phrasalen Strukturen ausgegangen: In (18a) und (18b) liegt keine Adjazenz vor, außerdem treten die fraglichen Konstituenten entweder als Kern einer Präpositionalphrase auf, vgl. *mit* in (18a) sowie Abb. 5, oder sie werden durch weitere Phrasen modifiziert wie *fort* in (18b). In (18c) stehen *leid* und *thun* zwar adjazent, *leid* wird jedoch durch das Indefinitpronomen *kein* spezifiziert, weswegen man in diesem Fall von einer vollständigen Nominalphrase ausgehen muss.

(18) a. [...] einige Stücke **mit** sich führete. (M 372)
b. [...] sich [...] **fort** nach Niemiecrowice begeben [...] (M 2387)
c. [...] niemandt kein **leid** thun [...] (AC 294)

Wie bereits bei den Kompositafällen können auch hier etwaige Zweifelsfälle mit TIGERSearch gefunden werden, indem man nach Wörtern mit dem Wortartlabel PTKVZ sucht, die adjazent zu einem Verb stehen.[10]

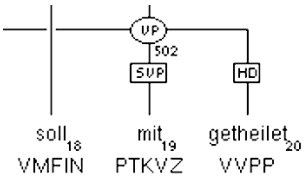


Abb. 4 Morphologische Struktur

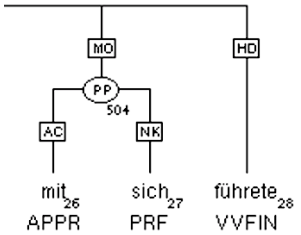


Abb. 5 Syntaktische Struktur

Am Anfang dieses Abschnitts wurde die gewählte Vorgehensweise, Partikelverben im Frühneuhochdeutschen unter anderem vor dem Hintergrund gegenwartsdeutscher Sprachkompetenz zu annotieren, angesprochen. Allerdings finden sich in älteren Texten des Deutschen auch Konstruktionen, die gemäß der Kriterien (i) und (ii) als Partikelverben zu analysieren sind, ohne eine Entsprechung im Gegenwartsdeutschen zu haben (19), wofür es unterschiedliche Gründe geben kann: So kennt man aus dem Gegenwartsdeutschen das Partikelverb *vortragen*, nicht aber *für tragen* (19a). Ebenso werden in älteren Texten Lexeme verwendet, die im Gegenwartsdeutschen fehlen, wie z.B. *anhero* in (19b). Und schließlich können auch komplexe Verben vorkommen, bei denen wir das Erstglied als solches aus dem Gegenwartsdeutschen kennen, die Kombination aus dieser Partikel mit dem darauffolgenden Verb im Gegenwartsdeutschen jedoch kein usualisiertes Wort mehr ist, wie z.B. im Fall von *nieder säbeln* in (19c).[11]

(19) a. [...] wie [...] denselben Ständen **für getragen** worden. (AC 15)
b. [...] etliche der ihrigen Officirer **anhero zu schicken** [...] (M 5909)
c. [...] auch andere [...] **nieder säbelten**. (M 370)

Für derartige Fälle wird von uns vorgeschlagen, diese als Partikelverben zu annotieren, falls sich ein Bezug zu einem Partikelverb im Gegenwartsdeutschen herstellen lässt. So wird *für tragen* in (19a) als Partikelverb analysiert, da es im Sinne vom Gegenwartsdeutschen *vortragen* gebraucht wird. Auch *anhero schicken* in (19b) wird als Partikelverb analysiert, weil es hier in der Bedeutung von 'hierherschicken' gebraucht wird, das im Gegenwartsdeutschen als Partikelverb angesehen wird. Bei *nieder säbeln* kann man ebenfalls von einem Partikelverb im Sinne von gegenwartsdeutschen Partikelverben *niederschlagen* bzw. *niederstechen* ausgehen. Weitere ähnliche Zweifelsfälle sind in (20) aufgeführt:

(20) a. [...] derohalben sie außtruckenlichen beuelch haben / höchst gedachten Römischen Keyser / von dieser Fürsten wegen / zuermahnen / daß er / die erwünschte gelegenheit nit **fürüber gehn** lasse / [...] (AC 579)
b. Von Cerigo wird berichtet/ daß die 40. Türckische Galeen / so den Succurs

in Candia gebracht / sich bereits wiederumb zu rücke begeben hätten mehr grosse Häupter und Völcker abzuholen und **über zuführen**.[12] (M 423)

Dagegen werden die Fälle unter (21) nicht als Partikelverben annotiert, da sie sich im Gegenwartsdeutschen nicht mit Partikelverben, sondern mit syntaktischen Strukturen in Verbindung bringen lassen: So ist aus dem Kontext klar ersichtlich, dass *dahinden geblieben* in (21a) sich auf die syntaktische Struktur *dahinten/hinten geblieben* und *dahinden lassen* in (21b) sich ebenfalls auf eine solche Struktur *dahinten/hinten lassen* zurückführen lassen und nicht mit den formal sehr ähnlichen gegenwartsdeutschen Partikelpräfixverben *hinterbleiben* und *hinterlassen* zu verwechseln sind. Solche Fälle werden folglich als syntaktische Strukturen annotiert.

(21) a. Die vnsere seind zu Gran / den 25. Aprilis / glücklich wider ankommen / deren mit mehr als 8. **dahinden geblieben** [...] (AC 288)
b. [...] hat doch nach langem Scharmützlen / mit schaden der seinigen weichen / vnd seinen Rennfahnen **dahinden lassen** müssen. (AC 496)

Unter der dargestellten Vorgehensweise lassen sich nahezu alle Fälle erfassen. Nichtsdestotrotz gibt es im Kontext der Partikelverben noch einige Konstruktionen, deren Annotation problematisch bleibt. Auf diese Konstruktionen soll im Folgenden eingegangen werden.

3.2.2 Annotation von Partikelverben: Sonderfälle

Doppelpartikel oder Adverb? Aufgrund der bereits formulierten Kriterien könnte man in Fällen wie (22) bei *mit* davon ausgehen, dass es mit der anderen Verbpartikel eine Art Doppelpartikel bilden könnte, wie etwa das Verb *miteinbeziehen* im Gegenwartsdeutschen.

(22) a. Die Römisch-Catholische Städte sind zwar von den Protestantischen sehr ersuchet worden / sich dieser Sachen **mit anzunehmen** [...] (M 1690)
b. [...] Etliche haben im abziehen ihrer Hanschke das Fleisch von allen Fingern **mit abgezogen**. (M 1715)

Unseres Erachtens ist jedoch die Analyse von *mit* als Adverb im Sinne von 'auch', 'ebenfalls' in diesen Fällen aus folgenden Gründen vorzuziehen: Zum einen scheint das Vorkommen von *mit* als Adverb im Deutschen – im älteren (23a) wie im heutigen (23b) – im Gegensatz zu anderen Präpositionen ziemlich gebräuchlich zu sein (mehr dazu vgl. auch die Wörterbucheinträge zum Lexem *mit* im Deutschen Wörterbuch von Jakob und Wilhelm Grimm (1999) für das ältere und im Duden-Wörterbuch (2007) für das heutige Deutsch):

(23) a. [...] **mit** nach Genua gezogen [...] (AC 364)
b. Er wollte **mit** nach oben gehen/ **mit** Tee trinken/ sich **mit** auf die Prüfung vorbereiten.

Im Gegensatz dazu kommen im Gegenwartsdeutschen kaum Doppelpartikeln mit *mit* vor,[13] schon gar nicht solche, die aus *mit* und einer weiteren Verbpartikel mit einfachem oder komplexem Adverb als Erstglied zusammengesetzt werden, wie die Beispiele aus dem Korpus in (24) zeigen. Das spricht unserer Meinung nach gegen die Lexikalisierung dieses Sprachmusters im Deutschen, weswegen wir uns in diesen Fällen für eine Annotation von *mit* als Adverb (s. Abb. 6) entschieden haben:

(24) a. [...] 30. der reichsten Kauffleute [...] **mit** weggeführet [...] (M 6794)
b. [...] eine schöne Koppel Türkischer Pferde **mit** über nehmen [...][14] (M 5992)
c. [...] viel 1000. Seelen nebest dem Herrn Michowsky **mit** hinweg geführet [...] (M 401)

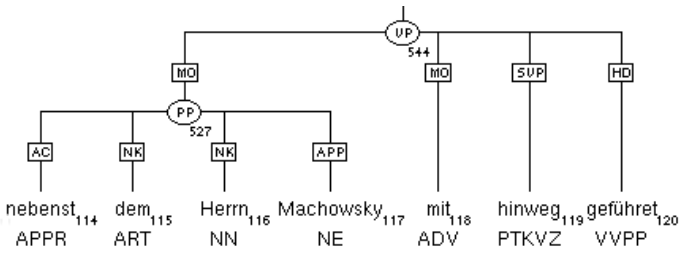


Abb. 6 *mit* als Adverb

Verbpartikel oder Adverb? Verben in Verbindung mit *da*-Adverbien wie *dahin* sollen hier insofern als besondere Gruppe hervorgehoben werden, als dass diese Adverbien ebenfalls im Gegenwartsdeutschen im Zusammenhang mit bestimmten Verben – entweder als Verbpartikel (25a) oder als Richtungsadverb (25b) – auftreten können:

(25) a. dahindämmern
b. dahin bringen

Der Unterschied zwischen morphologischer und syntaktischer Struktur im Gegenwartsdeutschen wird bei diesem Typ semantisch vorgenommen: Bei einer Konstruktion wie *dahin* + Verb wird dann von einer syntaktischen Struktur ausgegangen, wenn *dahin* in dieser Verbindung eine transparente Bedeutung besitzt, d.h. im Sinne von 'an diesen Ort', 'so weit' gebraucht wird (26a). Ansonsten ist bei *dahin* von einer Verbpartikel aus-

zugehen (26b). In diesem Fall weist die gesamte Konstruktion meistens eine lexikalisierte Bedeutung auf (Rat für deutsche Rechtschreibung 2006).

(26) a. Wir können zu Fuß **dahin gehen** (im Sinne von 'an diesen Ort gehen')
b. Die Tage sind **dahingegangen** (im Sinne von 'verstreichen')

Was die Analyse und Annotation im Frühneuhochdeutschen angeht, so soll hier das gleiche Kriterium angewendet werden. D.h. bei der Bestimmung der *da*-Wörter entscheidet die Bedeutung: Bildet *dahin* zusammen mit dem Verb eine gemeinsame lexikalisierte Bedeutung, dann wird es als Verbpartikel (PTKVZ) annotiert (27). Wenn *dahin* jedoch eine transparente Bedeutung im Sinne von 'an diesen Ort' besitzt, wird es als Adverb (ADV) annotiert (28).

(27) **morphologische Struktur**
Alle Consilia allhier gehen noch **dahin**$_{\text{PTKVZ}}$ [...] (M 434)

(28) **syntaktische Struktur**
Man hat nun vor/ den Handel auch mit denen von Algiers/ Tripoli/ Tunis und Biserta zu stabiliren/ und einen Ambassadeur mit etlichen Schiffen **dahin**$_{\text{ADV}}$ zu senden [...] (M 673)

Verbpartikel oder Präpositionalphrase? Dieser Abschnitt beschäftigt sich mit dreigliedrigen Strukturen im Frühneuhochdeutschen, wie sie in (29) zu sehen sind:

(29) a. Bald darnach haben die Kriegsleut [...] / 20. Türcken nidergehawet / vnnd 13. gefangen / vnnd solche **zu ruck** gebracht. (AC 954)
b. DJe Spanisch Armada, [...] ist auff den Spanischen Gränitzen angelangt / vnd durch vngewitter [...] **zu Grund** gangen. (AC 1059)

Aufgrund der Verbindung aus Präpositionalphrase und Verb könnte man diese Konstruktionen den syntaktischen Strukturen zuordnen. Aus dem Gegenwartsdeutschen weiß man jedoch, dass die Strukturen in (29) unterschiedliche Wege eingeschlagen haben: Die Verben mit adjazentem *zu ruck* (29a) haben sich durch Univerbierung zu den Partikelverben *zurückbringen, zurückgehen, zurückfahren* usw. entwickelt. Bei Konstruktionen wie in (29b)[15] handelt es sich dagegen nicht um Partikelverben, da hier der Prozess der Univerbierung noch nicht abgeschlossen zu sein scheint: Das zeigt sich beispielweise darin, dass diese Konstruktionen im Gegenwartsdeutschen immer noch als syntaktische Strukturen analysiert werden, die aus einem Verb und einem adverbialen Modifikator bestehen. Lediglich bei den ersten beiden Gliedern sind zwei verschiedene Analysen – als Adverb oder als Präpositionalphrase – erlaubt:[16]

(30) a. zugrunde$_{\text{ADV}}$/ [$_{\text{PP}}$ [zu$_{\text{APPR}}$] [Grunde$_{\text{NN}}$]] gehen *vs.* *zugrundegehen
b. zuwege$_{\text{ADV}}$/ [$_{\text{PP}}$ [zu$_{\text{APPR}}$] [Wege$_{\text{NN}}$]] bringen *vs.* *zuwegebringen

Vor diesem Hintergrund sollen hier unterschiedliche Analysen für die beiden Strukturen unter (29) angenommen werden: Die Fälle in (29a) werden als Partikelverben und somit

als morphologische Strukturen analysiert (s. dazu Abb. 7). Dafür spricht auch – freilich als eher schwächeres Argument –, dass man bei Fällen wie (29a) neben dem getrennten *zu ruck* auch das zusammengeschriebene *zuruck* (31) belegt findet:

(31) [...] die gelegenheit der Bawren aber ware jhme vnbekannt / derowegen er wider **zuruck gezogen**. (AC 296)

Konstruktionen wie in (29b) werden hingegen als syntaktische Strukturen annotiert. Eine vorgenommene Suche nach ähnlichen Fällen im Korpus (diese ist bei solch wenigen und speziellen Fällen wie in (31) im Gegensatz zum gesamten Bereich der Partikelverben relativ schnell zu leisten) führt zu dem Ergebnis, dass solche Konstruktionen in unserem Korpus immer mit einem Spatium aufgetreten sind, was dafür spricht, die Erstglieder als Präpositionalphrasen (32a) (s. dazu Abb. 8) und nicht als Adverb-Komposita (32b) zu analysieren.

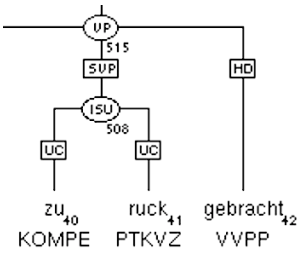


Abb. 7 Partikelverb

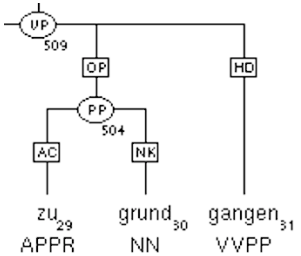


Abb. 8 Präpositionalphrase

(32) a. [PP [zuAPPR] [GrundeNN]] gehen
b. *[AVP [zuKOMPE] [GrundeADV]] gehen

4 Zusammenfassung

Wir haben im vorliegenden Artikel das MERCURIUS-Korpus vorgestellt, welches bislang aus zwei mit syntaktischen Informationen angereicherten frühneuhochdeutschen Zeitungstexten besteht. Um die spezifischen Probleme aufzuzeigen, die sich bei der syntaktischen Annotation historischer Korpora ergeben, haben wir uns im vorliegenden Beitrag mit sprachlichen Strukturen beschäftigt, denen sowohl Wortstatus als auch phrasaler Status zugeschrieben werden kann. Für Partikelverben und (andere) Kompositionsbildungen wurde gezeigt, wie solche Muster in einem zeitlich vertretbaren Rahmen dennoch deskriptiv adäquat annotiert werden können.

Anmerkungen

[1]Wie das händisch bereits in den Arbeiten geschehen ist, die in der Reihe ›Bausteine zur Sprachgeschichte des Neuhochdeutschen‹ ab 1964 erschienen sind.

[2]Näheres zur Digitalisierung im Allgemeinen und der Segmentierung der Satzgrenzen im Speziellen findet sich in Demske et al. (2004) und Demske (2007).

[3]Diese als nicht-parallele Koordinationsellipse bekannte Konstruktion ist im Frühneuhochdeutschen sehr gebräuchlich.

[4]Mit TIGERsearch als Suchwerkzeug kann eine Suche nicht nur nach einzelnen Lexemen, Wortarten oder Phrasen getätigt werden, sondern es können auch spezielle und explizite Suchanfragen ausgeführt werden, wie z.B. die Suche nach einer Nominalphrase, die ein vorangestelltes Genitivattribut enthält, oder nach einem Satz, der einen Komplementsatz enthält.

[5]Wir werden bei den Korpusbeispielen im Folgenden nur dann einen Kontext mitliefern, wenn ein solcher erforderlich ist, um die entsprechende Konstruktion verstehen zu können. Ansonsten werden wir – mit Ausnahme des zu Illustrationszwecken gezeigten Satzes in (7) – stets nur die fragliche Struktur selbst abbilden.

[6]Angegeben ist jeweils nach der Abkürzung für den spezifischen Text die Satznummer im elektronischen Korpus.

[7]Mehr zu diesen und ähnlichen Unterscheidungskriterien vgl. Rat für deutsche Rechtschreibung (2006) und Zur Neuregelung der deutschen Rechtschreibung (2011).

[8]Die orthografische Regelung ist hierbei selbstverständlich nur als Reflex der linguistischen Verhältnisse zu sehen. Dies heißt, dass selbst ein fälschlicherweise mit Spatium versehenes *kaltstellen* trotzdem aufgrund linguistischer Kriterien wie der semantischen Verschiebung ein Partikelverb bleibt, und zwar unabhängig von der Schreibung.

[9]Gemeint sind hier die Fälle, in denen Verbpartikel und Verb topologisch gesehen zusammen in der rechten Satzklammer auftreten.

[10]Zu Näherem bzgl. einer ähnlichen Suchanfrage siehe Müller (2006).

[11]So kommt *niedersäbeln* bspw. im mehrbändigen duden-Wörterbuch (1999) mit mehr als 200.000 Stichwörtern gar nicht mehr vor.

[12]Das Verb *über führen* wird hier im Sinne vom gegenwartsdeutschen Partikelverb *herüberführen* und nicht dem Partikelpräfixverb *überführen* gebraucht.

[13]Bei der Recherche in Online-Duden (www.duden.de) stößt man nur auf wenige Doppelpartikelverben mit *mit* wie bspw. *miteinbeziehen* oder *miteinrechnen*.

[14] *Über* wird hier im Sinne von 'herüber' verwendet.

[15]Ähnliche Fälle wären *zu Wege/zuwege, zu Stande/zustande, zu Lasten/zulasten bringen*.

[16]Die Label in den nachfolgenden Beispielen wurden aus Annotate übernommen: PP steht für 'Präpositionalphrase', APPR für 'Präposition', NN für 'Nomen', AVP für 'Adverbphrase', KOMPE für 'Kompositions-Erstglied' und ADV für 'Adverb'.

Quellen

[*M*] = Mercurius 1667. Nordischer Mercurius. Welcher kürzlich vorstellet/was in diesem 1667. Jahre an Novellen aus Europa einkommen ist. Hamburg 1667.

[*AC*] = Annus Christi 1597. Historische erzöhlung/der fürnembsten Geschichten vnd handlungen/so in diesem 1597. Jahr (...) abgelaufen (...). Rorschach 1597. Nachdruck: Walluf-Nedeln: Sändig 1977.

Literatur

Admoni, V. G. (1980). Zur Ausbildung der Norm der deutschen Literatursprache im Bereich des neuhochdeutschen Satzgefüges (1470-1730). Ein Beitrag zur Geschichte des Gestaltungssystems der deutschen Sprache. Akademie-Verlag, Berlin.

Bech, K. und Eide, K. G. (2011). The annotation of morphology, syntax and information structure in a multilayered diachronic corpus. JLCL, 26(2):13–24.

Brants, S. (1999). Tagging and Parsing with Cascaded Markov Models – Automation of Corpus Annotation. DFKI, Saarbrücken Dissertations in Computational Linguistics and Language Technology Bd. 6.

Brants, T. (2000). Inter-annotator agreement for a German newspaper corpus. In Proceedings of Second International Conference on Language Resources and Evaluation LREC-2000. Athen.

Brants, S., Dipper, S., Hansen, S., Lezius, W., und Smith, G. (2002). The TIGER treebank. In Proceedings of the Workshop on Treebanks and Linguistic Theories, 24–41, Sozopol.

Brants, T., Wojciech, S., und Uszkoreit, H. (1999). Syntactic annotation of a german newspaper corpus. In Proceedings of the ATALA Treebank Workshop, 69–76, Paris.

Demske, U. (2001). Merkmale und Relationen: Diachrone Studien zur Nominalphrase des Deutschen. de Gruyter, New York, Berlin.

Demske, U. (2007). Das Mercurius-Korpus: Eine Baumbank für das Frühneuhochdeutsche. In Kallmeyer, W. und Zifonun, G., Hgg., Sprachkorpora – Datenmengen und Erkenntnisfortschritt, 91–104. de Gruyter, Berlin, New York.

Demske, U., Frank, N., Laufer, S., und Stiemer, H. (2004). Syntactic interpretation of an early new high german corpus. In Kübler, S., Nivre, J., Hinrichs, E., und Wunsch, H., Hgg., Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004), 175–182, Tübingen.

Deutsches Wörterbuch von Jacob und Wilhelm Grimm (1999). Nachdruck der Erstausgabe von 1854-1961. Deutscher Taschenbuchverlag, München.

Duden (1999). Duden – Das große Wörterbuch der deutschen Sprache in 10 Bänden. 3., völlig neu bearbeitete und erweiterte Auflage. Band 6. Dudenverlag, Mannheim, Leipzig, Wien, Zürich.

Duden (2007). Duden – Deutsches Universalwörterbuch. 6., überarbeitete und erweiterte Auflage. Dudenverlag, Mannheim, Wien, Zürich.

Duden (2009). Duden – Die Grammatik. 8., überarbeitete Auflage. Dudenverlag, Mannheim, Wien, Zürich.

Eisenberg, P. (2006). Grundriss der deutschen Grammatik. Band 1: Das Wort. 3. Auflage. Metzler, Stuttgart, Weimar.

Helbig, G. and Buscha, J. (2007). Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. 6. Auflage. Langenscheidt, Berlin, München, Wien.

Kroch, A. und Taylor, A. (2000). Penn-Helsinki Parsed Corpus of Middle English.

Lezius, W. (2002). TIGERSearch – Ein Suchwerkzeug für Baumbanken. In Busemann, S., Hg., Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache, Saarbrücken.

Müller, S. (2006). Quantitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus. In Kallmeyer, W. und Zifonun, G., Hgg, Sprachkorpora – Datenmengen und Erkenntnisfortschritt, 70–90. de Gruyter, Berlin, New York.

Ramers, K. H. (1997). Die Kunst der Fuge: Zum morphologischen Status von Verbindungselementen in Nominalkomposita. In Dürscheid, C., Schwarz, M., und Ramers, K. H., Hgg., Sprache im Fokus. Festschrift für Heinz Vater zum 65. Geburtstag, 33–46. Niemeyer, Tübingen.

Rat für deutsche Rechtschreibung (2006). Deutsche Rechtschreibung. Regeln und Wörterverzeichnis. Entsprechend den Empfehlungen des Rats für deutsche Rechtschreibung. Überarbeitete Fassung des amtlichen Regelwerks mit den Nachträgen aus dem Bericht 2010.

Smith, G. und Eisenberg, P. (2000). Kommentare zur Verwendung des STTS im NEGRA-Korpus. Manuskript.

Voormann, H. und Lezius, W. (2002). TIGERin – Grafische Eingabe von Benutzeranfragen für ein Baumbank-Anfragewerkzeug. In Busemann, S., Hg., Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache.

Zur Neuregelung der deutschen Rechtschreibung ab 1. August 2006 – Nachtrag (2011). Sprachreport. Extra-Ausgabe. Institut für Deutsche Sprache, Mannheim.

Jolanta Gelumbeckaitė, Mindaugas Šinkūnas, Vytautas Zinkevičius

Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation

1 AIMS

The *Old Lithuanian Reference Corpus* (Lith. *Senosios lietuvių kalbos korpusas*; acronym SLIEKKAS, cf. Lith. *sliekas* "earthworm") is a comprehensive, deeply annotated diachronic reference corpus of Old Lithuanian, being developed in cooperation between the Goethe-University of Frankfurt/Main (Germany), the Institute of Lithuanian Language (Vilnius, Lithuania), and the University of Pisa (Italy)[1]. The aim of the project is to create a multimodal (alignment of the annotated texts with facsimile reproductions of the original), annotated (header-information, hierarchic, structural, palaeographic, textological, lexical, and grammatical annotations) reference corpus (meta-linguistic information about Old Lithuanian, its diatopic variations, characteristic vocabulary). The ultimate goal is to develop a qualitative multilevel electronic retrieval engine for multilateral linguistic research of Old Lithuanian which will lead to reliable results for diachronic Lithuanian language studies. It will enable the implementation of the two biggest desiderata of Baltic linguistics: the Old Lithuanian grammar, and the historic dictionary of Lithuanian.

The most suitable technological and scientific basis for the multi-layer stand-off annotations is to be established on the basis of 10 selected texts (cf. Section 2): lemmatising (main word form and attested word form, the latter both in a transliterated form and as a normalised form in Modern Lithuanian), glossing (standard form of the lemma and of the attested word as well as their meanings), hierarchic grammatical description, predominantly restricted to morphology (part-of-speech tagging, flexional morphology of the lemmata and single attested word forms), and alignment of the annotated Lithuanian texts with each other and with their Polish, Latin, German etc. translation source texts.

The main endeavours of SLIEKKAS are the following: 1) securing a high philological standard as well as a textological and a palaeographic annotation of the selected Old Lithuanian texts, 2) setting up a basic-XML-structure, which is relevant for a further annotation, and 3) digitisation of the Lithuanian lexica and word indices, which are relevant for a further lemmatising and glossing of the texts.

The *Old Lithuanian Reference Corpus* is designed to provide an innovative scientific resource for historical and comparative linguistics as well as literary, religious, and cultural studies of the Baltic countries. This also includes materials related to the controversy between pre-Christian and Christian cultures and the confessional spin-off processes of the area as well as their backgrounds. In this way, the essential knowledge of the cultural development of Lithuania and the Baltic countries in the given period will be gained. With regard to historical linguistics, the *Old Lithuanian Reference Corpus* is expected to provide a basis for an efficient development and implementation of further research programmes concerning the diachronic grammar and the lexicon of Lithuanian.

This paper focuses on the main steps towards a semi-automated human-controlled grammatical annotation of Old Lithuanian, the available resources, and the most suitable software for this purpose.

2 MATERIAL

Old Lithuanian covers a period of ca. 300 years, from the 16th to the 19th centuries. The earliest known coherent Lithuanian text consists of three so called "Dzūkian prayers" in the copy of *Tractatus sacerdotalis* by NICOLAUS DE BLONY, preserved at the Vilnius University library (Straßburg: Martin Flach; Sign.: VUB RS II–3006). The year 1800, with the grammar by CHRISTIAN GOTTLIEB MIELCKE (1732–1807) *Anfangs-Gründe einer Littauischen Sprach-Lehre* (Königsberg: Hartung), marks the beginning of the standardisation and codification of Lithuanian based on a more or less single dialect, i.e. the southern group of the West High Lithuanian (=West Aukštaitian) dialect.

In total, the corpus will consist of over 10 million text words. Due to such a huge amount and to the complex, multilayered structures, which are needed for such a diachronic corpus, it seems reasonable to start with a smaller test corpus. Ten Old Lithuanian texts comprising ca. 350 000 tokens were chosen for this test corpus:

1. *DzP* ca. 1520—"Dzūkian prayers" (consisting of *Pater noster, Ave Maria, Credo*), the oldest known Lithuanian text; manuscript; translation from Latin, Polish, and/or German.
2. *MžK* 1547—MARTYNAS MAŽVYDAS, *Catechismusa prasty ßadei*; the oldest printed Lithuanian book; partly translated from Latin and Polish, and partly original written text.
3. *MžGA* 1549—MARTYNAS MAŽVYDAS, *Giesme S. Ambraßeijaus*; print; partly translated from Latin and Polish, and partly original written text.
4. *MžFK* 1559—MARTYNAS MAŽVYDAS, *Forma Chrikstima*; print; translation from German.
5. *WP* 1573— *Wolfenbüttel Postil*; manuscript; partly translated from Latin and partly original written text.
6. *VE* 1579—BALTRAMIEJUS VILENTAS, *Enchiridion*; print; partly translated from Latin and German, and partly original written text.
7. *DK* 1595—MIKALOJUS DAUKŠA, *Kathechismas*; print; translation from Polish and Latin.
8. *LyK* 1719—HEINRICH JOHANN LYSIUS, *Mažas Katgismas*; manuscript; translation from German.

9. *EnK* 1722—GABRIEL ENGEL, *Mažas Katgismas*; print; elaborated version of Lysius' catechism.

10. *DM* ca. 1765/1775—KRISTIJONAS DONELAITIS, *Metai*; manuscript, the first Lithuanian poem, autochthonic text. Editions of the text: first edition by LUDWIG J. RHESA (*DMRh*1818); second edition by AUGUST SCHLEICHER (*DMSch*1865); third edition by GEORG H. F. NESSELMANN (*DMN* 1869).

The selected texts represent a characteristic variety of Old Lithuanian text genres, sorts and types—a) religious as well as secular texts, the religious texts being those of the prayers, catechisms, hymnals, and sermons, all of them including Bible quotations; b) prose and poetry, c) translated, original written, and compiled texts, d) translations from Latin, German and Polish, and e) handwritten as well as printed texts. The chosen texts stand for the three language variations of Old Lithuanian, determined according to their dialectal, sociolectal, and confessional features—the Western or so called Prussian (*VE*, *LyK*, *EnK*, *DM*), the Middle (*DK*), and the Eastern type (*DzP*) of Old Lithuanian as well as a compound of several dialects (*MžK*, *MžGA*, *MžFK*, *WP*).

The selected texts also differ in their spelling as well as their accentography, which documents different strategies in indicating a free word stress through the grave, acute, or circumflex accent-mark and in marking two types of syllable accents (for falling resp. rising tonemes) on the one hand, and which also belongs to the system of the diacritical marks (similar to the Neo-Latin practice of accentuation) on the other. Some texts are accentuated (*DK*, *DM*; partly *LyK*, *EnK*), others not. Being heterogeneous as such, the texts determine a rich representativeness of the test corpus by simultaneously causing additional problems for computer processing.

3 ARCHITECTURE

The intended annotation scheme of the Corpus embraces the following structural features:

1. *A thorough linguistic and textological annotation, including header information, lemmatisation, grammatical information (part-of-speech tagging, morphological and basic syntactical information), glossing (in Standard Lithuanian, English, and possibly other languages), information about the text structure (text subdivision into words, sentences, lines, verses, paragraphs, etc.), palaeographic (resp. typographic) and textological information*—

The main purpose is to develop a semi-automated technique that allows establishing the core word form in a historical lexicon (lemmatisation), its glossing in Standard Lithuanian and the determination of its actual meanings in a given Old Lithuanian text. More than 50% of the Old Lithuanian word forms are ambiguous as regards their morphological status. A morphological annotation consists of the unalterable morphological categories of the lemmata as well as of the actual word forms in a given text, and of the flexional morphological characteristics of the latter. For instance, the morphological categories of the lemma and of the attested word form in a

given text are to be annotated differently in such cases, as the masculine adjective *gražiausias* "the most beautiful", which belongs to the *ja*-paradigm (superlative form), while its lemma *gražus* (Masc), *graži* (Fem) belongs to the *u,jo*-paradigm—thus two separate levels for the morphological categories have to be created, one for the token, the other for its lemma. A distinction of the morphological categories of the lemma and of the actual word form helps to trace the alteration of grammatical classes in Old Lithuanian, e.g., substantivisation of adjectives, adjectivisation of participles, adverbalisation, etc. For example, the form *laukan* "to the outside, into the field" is a paradigmatic illative case of the substantive *laukas* "field" in Old Lithuanian, whereas the form *laukan* is considered merely as the adverb "out" in Standard Lithuanian.

These annotation levels (lemmatisation, glossing, part-of-speech tagging, and morphological annotation) are carried out on the basis of the *Toolbox* program (SIL: http://www.sil.org/computing/toolbox/). Afterwards, they will be revised and corrected in the annotation software *ELAN* (Max Planck Institutefor Psycholinguistics in Nijmegen, http://tla.mpi.nl/tools/tla-tools/elan/). Furthermore, the texts will be provided with the basic information on the syntactic structure of the sentences (simple and complex sentences will be marked) in *ELAN* directly.

Single Latin, German, or Polish words and sentences within the Lithuanian texts will be annotated according to the morphology of a corresponding language. Additional annotation levels are required for the taxonomy of both explicit and implicit quotations in the Old Lithuanian texts. It enables a clear distinction between the translated resp. re-narrated text parts and the original written text.

2. *A multi-level architecture of the annotations*—

The aim is to generate an XML-structure that comprises all the intended annotation levels. The experience of the DFG project *Referenzkorpus Altdeutsch*[2] has shown that the software *ELAN* fully serves this purpose. The *ELAN* data structures can either be produced directly from the text data on the basis of the *Toolbox* program, or they can be generated from autonomously programmed components which incorporate the text data with their lexical, grammatical, and other information.

3. *Multi-modality of the corpus through the alignment of the texts with facsimile reproductions of the original*—

The Old Lithuanian texts will be aligned automatically with facsimile reproductions of the originals (manuscripts resp. prints) on a line level and additionally aligned manually on a word level.

Since most of the Old Lithuanian texts are translations from Latin, German, or Polish sources, the source texts (ca. 190 000 text words in the case of the test corpus) will be annotated in the same way as the Lithuanian ones. This will enable the alignment of the Old Lithuanian texts with their sources with respect to all annotation levels. Furthermore, the Old Lithuanian texts of the same genre will be aligned with each other in order to allow for an assessment of possible mutual influences within a single genre as well as across genres.

4 RESOURCES FOR ANNOTATION

Old Lithuanian can be roughly classified into three main periods of the evolution of orthography. The early period is the most variable and unstable one. Orthography gets more uniform in the middle of the 17th century in Lithuania Minor (Duchy of Prussia), but it has a different variant in Lithuania Major (Grand Duchy of Lithuania). The specific orthography of the texts has to be converted (during which the regular dialectal phonetic features are discarded) to match the one that exists in Modern Lithuanian, in order to be processed by an automatic morphology analyser. Results of these processes will be included in the annotation levels of the "Standardised word form (transliteration)" and "Normalised word form (in Modern Lithuanian)", according to which retrieval tasks can be modified. The conversion of the old orthography to the modern one is done by the transliteration rules that are implemented using the *Consistent Changes Program* (SIL: http://www.sil.org/computing/catalog/show_software.asp?id=4). For the orthography of the early period, special rules have to be created for every individual author (sometimes even every text of the same author). The (ortho)graphy of the texts from the 16th century differs from the one used in the 18th century (cf. Ill. 1 and 2). The transliteration rules are more stable for the later period, though they are also slightly modified for each author (or text) to attain the maximum possible accuracy.

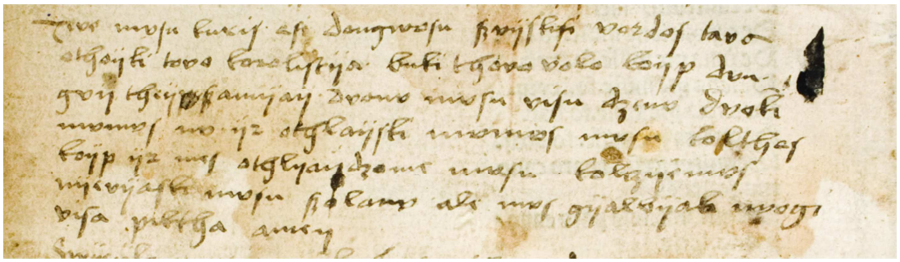


Illustration 1: A fragment of *DzP*, ca. 1520 (*Pater noster*)

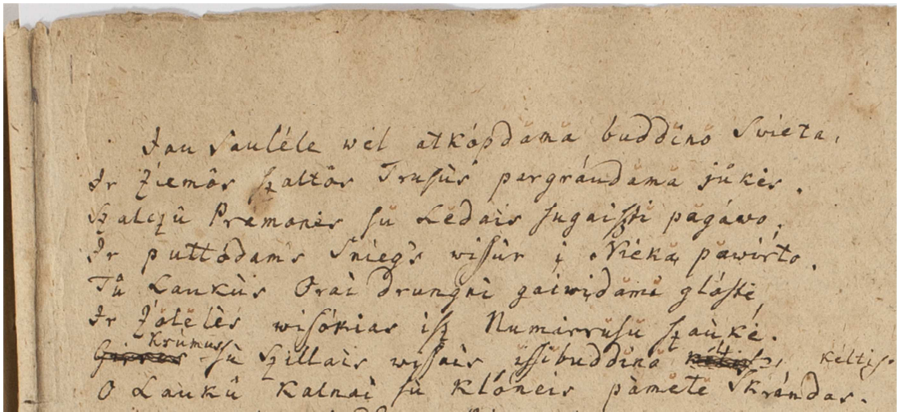
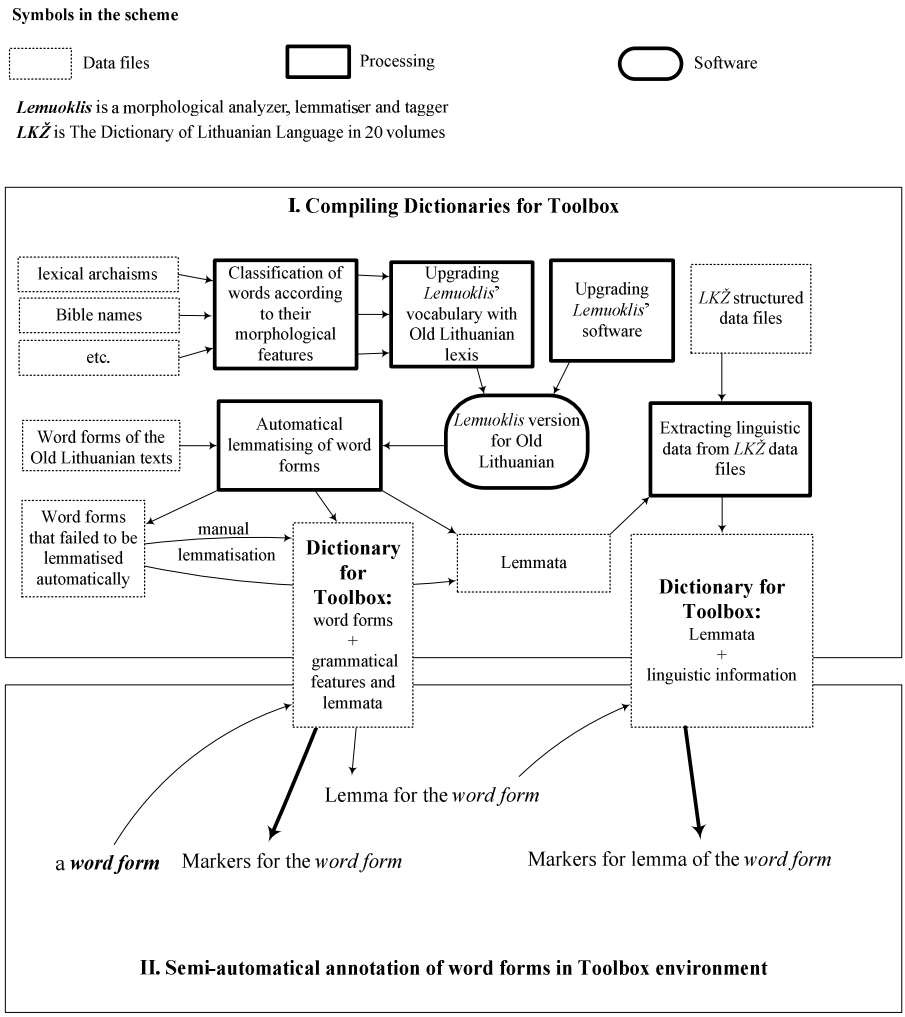


Illustration 2:A fragment of *DM*, ca. 1765/1775

To give an example of the transliteration, in the words *Ape Swetaſti* "about the Sacrament" (chapter name, *MžK* 25) the rules of changing the long <ſ> to the round <s> and <w> to <v> in the form *Swetaſti* are applied, and a form *svetasti* is created. The original word form layer with *Swetaſti* remains unchanged. By implementing additional rules the created form *svetasti* gives possible correspondences, namely *svetastį*, *svėtasti*, and *svėtastį*. Only the latter form will be recognised as a valid entry with standardised orthography in the Old Lithuanian database and can then be analysed further.

Figure 1 shows resources and processes used for developing a semi-automated technique for the grammatical annotation of the Old Lithuanian words. In the annotation process, the *Toolbox* environment utilises two dictionaries, as is shown in the lower part of the scheme. Firstly, a search for a word form in the dictionary of the word forms is performed: lemma, part-of-speech, and other grammatical markers for the word form are extracted. Afterwards, markers for the lemma are searched and extracted from the dictionary of the lemmata. In case the search results are ambiguous, i.e., when more than one record is found in a dictionary, the annotator working with the *Toolbox* program must make a decision and choose the correct variant. In order to enable these processes, two dictionaries—one of the Old Lithuanian word forms and another of the Old Lithuanian lemmata—are currently being compiled, as is shown in the upper part of Figure 1.

Figure 1. The Automated Grammatical Annotation in SLIEKKAS



While producing the Old Lithuanian dictionary, the word forms are lemmatised and POS-tagged using the software *Lemuoklis*, a morphological analyser, lemmatiser, and tagger for Modern Lithuanian (ZINKEVIČIUS 2000, ZINKEVIČIUS, DROŻDŻYŃSKI, HOMOLA, PISKORSKI 2003). *Lemuoklis* is a rule-based system. The lexical and grammatical data of the program

consist of several lexica (organised as letter trees). Three of them store the roots of Lithuanian words, which are associated with certain appropriate morphological rules; morphological rules are presented in the form of digital tables. Both the vocabulary of stems (organized as a tree data structure) and tables of rules are in the original internal digital format. *Lemuoklis* is a library of functions programmed using the C++ language. Other lexica store word forms with no morphological informationor contain lists of abbreviations and acronyms (http://donelaitis.vdu.lt/~vytas/tool/tool.ppt). The software is able both to analyse a word form grammatically and to synthesise a new inflectional form. It performs lemmatisation by means of synthesising new forms (e.g., nominatives for nouns and infinitives for verbs and verbal forms).

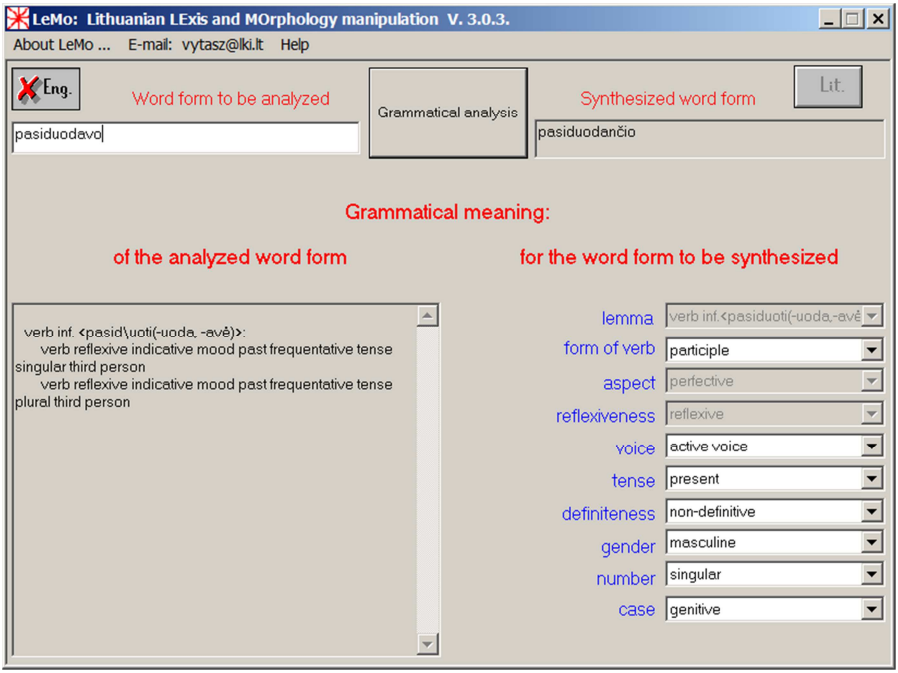


Figure 2: Demo window of *Lemuoklis* – a morphological analyser, lemmatiser, and tagger for Modern Lithuanian

Figure 2 shows a demo and testing window of *Lemuoklis.* The word form *pasiduodavo* (past frequentative, "was used to surrender") is analysed by *Lemuoklis* (in the left part of the window) as having the lemma *pasiduoti* (the infinitive, "to surrender") and is characterised with the tags "verb reflexive", "indicative mood", "past frequentative tense", "plural", and "third person". The infinitive form is supplied with the endings of the two other main forms of a verb, i.e., present and past forms, *pasiduoda* and *pasidavė* respectively. In the right part of the window the annotator can select morphological properties for a new inflectional form of the word *pasiduoti* ("to surrender") which is synthesised automatically in the upper box "Synthesised word form". In case a surface form is homonymous, i.e., it has several grammatical meanings, the program

gives full grammatical characteristics for each possible homograph of the surface form. However, some methods are used to reduce the ambiguity without taking into account the context: one of them is the method of disambiguation between diminutive nouns with the suffix *-yti(s)* and respective verbal infinitive forms. For example, the Lithuanian word form *padaryti* is interpreted as a transitive infinitive form ("to do something") rather than a theoretically possible voc. sg. form of a diminutive *padarytis* from *padaras* ("a creature"); the word form *ginčytis* is interpreted as a reflexive infinitive ("to argue") rather than the nom. sg. of a diminutive *ginčytis* from *ginčas* ("a dispute, argument"). The disambiguation between proper and common nouns is performed through the application of special lexica containing proper noun forms from Modern Lithuanian corpora and other sources (ZINKEVIČIUS, DROŽDŽYŃSKI, HOMOLA, PISKORSKI 2003).

The original *Lemuoklis* is based on the Modern Standard Lithuanian grammar and various modern lexica. In order to enable *Lemuoklis* to recognise words from the Old Lithuanian texts it was enriched through a special vocabulary which comprises the dictionaries of Old Lithuanian (PALIONIS 2004; ca. 8000 words) and of Slavic loanwords in Old Lithuanian (SKARDŽIUS 1998; 4152 words) as well as the dictionary of Bible names (KIMBRYS 2000; 3251 words), and some other lexical material. All added words had to be classified semi-manually (while choosing correct answers to the questions, cf. Fig. 3) according to their morphological features while using a special software, which creates supplemental lexica of the roots associated with morphological rules for *Lemuoklis*.

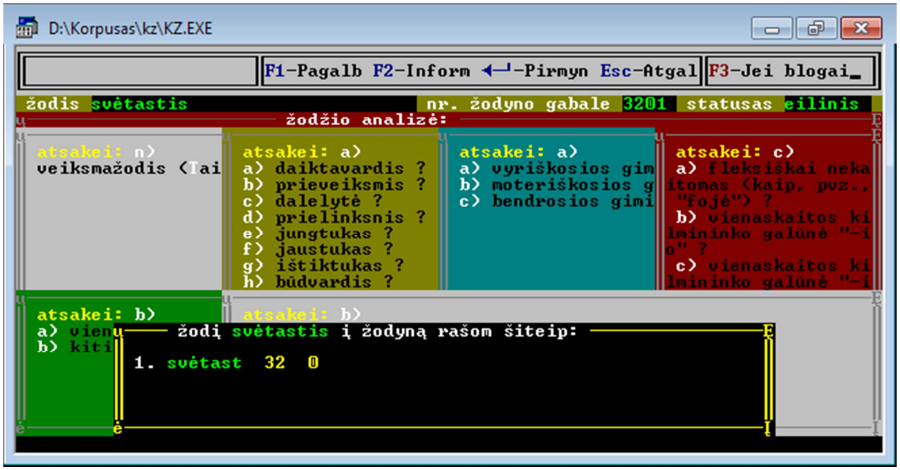


Figure 3: Process of semi-manual classification of words from the Old Lithuanian texts

Figure 3 shows the process of classifying the Old Lithuanian word *švėtastis*. During the first step, the system formulated the question "is it a verb?" and an operator answered by pressing "n" (no). In the second step, the system enumerated names of the parts of speech, and an operator's choice was "a noun" (*daiktavardis* in Lithuanian) by pressing "a"; during the next step, gender was defined (*vyriškoji* masculine, *moteriškoji* feminine, *bendroji* common). Then, an operator was asked whether *švėtastis* is a non-inflective (variant *a*) form, or having the genitive ending *-io*

(variant *b*) resp. *-ies* (variant *c*), and an operator chose the variant "c". Next, the possibility to build a plural form was confirmed, and a type of declension was specified more precisely by choosing the right ending variant for the plural genitive. The two lower windows in the screen were covered by the final black one which indicates the result of the word classification process: "svėtast 32 0", where *svėtast-* is the stem, *32* is an internal number for the inflection type, and *0* indicates the number of letters at stem's end that differs through the inflectional paradigm.

To get back to the above-mentioned example of the transliteration of *Swetaſti*, *Lemuoklis* is provided with the accusative forms *svetasti*, *svetastį*, *svėtasti*, and *svėtastį*. The word *svėtastis* (Nom) does not exist in Modern Lithuanian, but the root was added together with other loanwords from SKARDŽIUS' Old Lithuanian dictionary (SKARDŽIUS 1998), and thus can be processed by the modified version of *Lemuoklis*. In the process of an automatic lemmatisation and of an analysis of the word forms *svetasti, svetastį*, *svėtasti*, and *svėtastį,* only the latter form *svėtastį* is recognised by *Lemuoklis*, its flexional morphological characteristics (Sg_Acc), flexional class (i_Fem), part-of-speech (noun), and lemma (*svėtastis*) are generated. This information is stored in the SLIEKKAS dictionary of the word forms.

While producing the lemmata dictionary, the required grammatical information is obtained from the Lithuanian language dictionary (*LKŽ*; *Lietuvių kalbos žodynas*, 20 volumes, printed in 1941–2002; online: www.lkz.lt/startas.htm; ZINKEVIČIUS 2008). This thesaurus includes ca. ½ million lemmata. The words which are essential regarding the needs of the testcorpus have been extracted according to the token list of the test corpus. For instance, the lemma *svėtastis* was searched in *LKŽ* and marked with the following information: accented lemma (*svẽtastis*), part-of-speech (noun), and accentuation class (1); the flexional class is additionally created by the retrieval engine (i9_Fem). This information is stored in the SLIEKKAS lemmata dictionary.

5 DISAMBIGUATION

Three types of data are created by means of the software and lexical resources mentioned above: 1) a list of transliterated word forms, 2) a dictionary of normalised word forms, which includes information on the part-of-speech, unalterable morphological categories, and flexional morphological characteristics of the actual word form, and 3) a dictionary of the lemmata, which includes the tags for part-of-speech and unalterable morphological categories of the lemma, and also its accentual class. Separate dictionaries for the translation into other languages (English and possibly German) can be added while linking them to the Lithuanian lemmata. The abovementioned three types of data (for transliterated and normalised word forms as well as for lemmata) shall be managed by the *Toolbox* program, in which the annotation levels (lemmatisation, glossing, part-of-speech, and morphological annotation) are created and disambiguation is controlled by a human (Fig. 4).

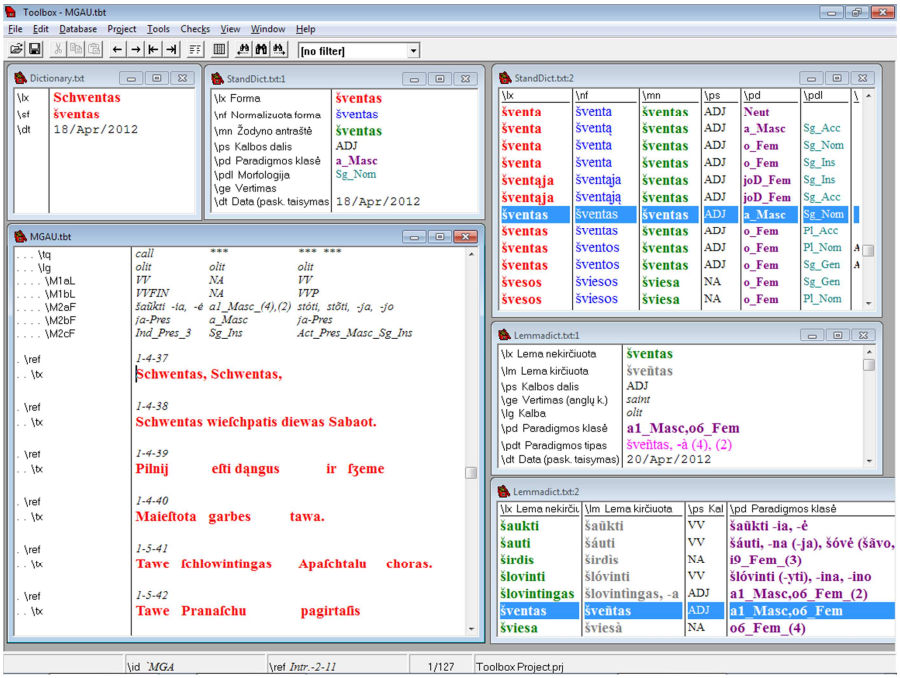


Figure 4: Annotations are created in the *Toolbox* program using the generated lexical and grammatical information

Figure 4 illustrates the *Toolbox* environment, where the word *Schwentas* ("saint") is being processed in the window MGAU.tbt using a list of transliterated forms (window Dictionary.txt), a dictionary of the normalised word forms (StandDict.txt) and a dictionary of the lemmata (Lemmadict.txt). The rich flexion of Lithuanian and the inconsistency of the old orthography result in a very high rate of homographs. The automatic disambiguation is complicated because the analysis is done on the word level only without involving the context or considering punctuation (no tools or rules on the Old Lithuanian syntax are implemented), in the absence of semantic information, without regard to accent marks, and with lack of statistical data. The overall disambiguation has to be controlled manually, as can be seen in Figure 5 (the word form *šventas* can be either Masc_Sg_Nom or Fem_Pl_Acc). After the ambiguous grammatical information is dissolved, the annotation layers are created (Fig. 6).

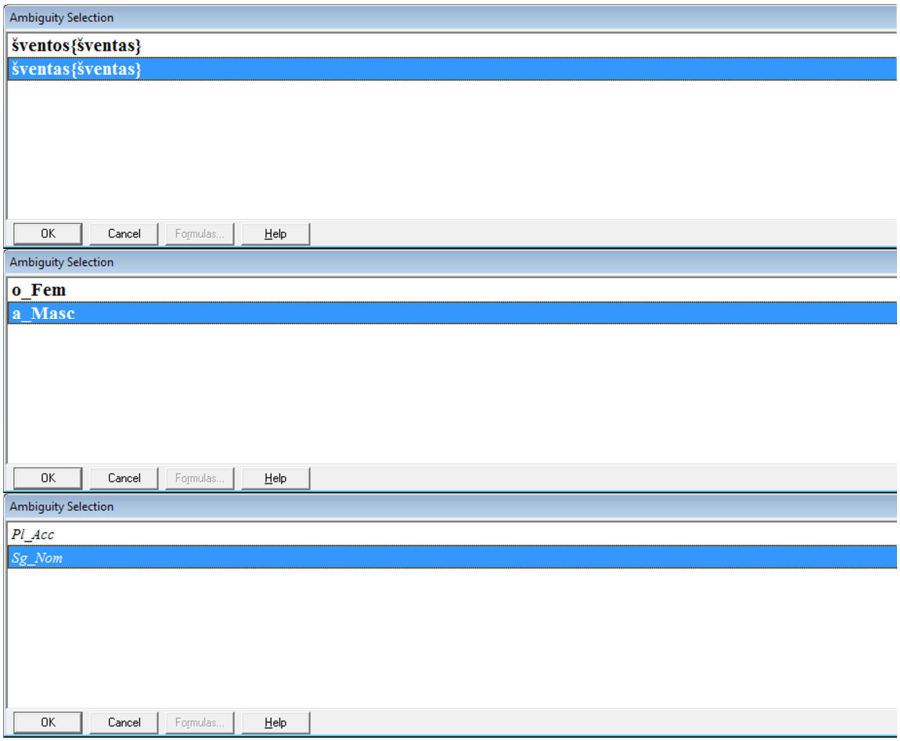


Figure 5: Steps of ambiguity selection for the annotation of the word form *Schwentas* "saint"

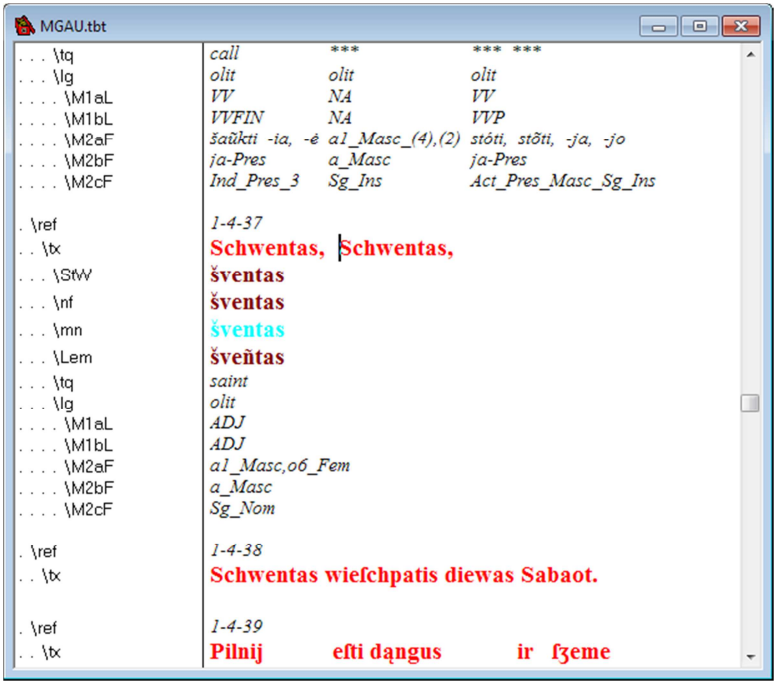


Figure 6: Steps of ambiguity selection for the annotation of the word form *Schwentas* "saint"

Afterwards, the annotations will be revised and corrected by an annotator in the software *ELAN*. Furthermore, the texts will be provided with the basic syntactic information in the software directly. Single Latin, German, or Polish words and sentences that occur within the Lithuanian texts will be annotated by hand correspondingly.

6 CONCLUSION

The multi-layer stand-off architecture (the architecture, in which every layer is a separate document, and nonetheless all layers are synchronised among themselves) of the tags and the amount of the texts in the test phase of the *Old Lithuanian Reference Corpus* require solutions for automated processes that could help to save time. This can be achieved using different databases, compiled from available lexical and grammatical resources. The morphological annotation of the Old Lithuanian word forms can be done using a modified version of the morphology analysis software of Modern Lithuanian. Nevertheless, the rich inflection of Lithuanian results in a very high rate of homographs. Their grammatical disambiguation still has to be solved manually.

[1]In 2010 SLIEKKAS was supported by the Lithuanian Ministry of Education and Science. Since 2012 it is funded by a grant No. VAT-42/2012 from the Research Council of Lithuania and performed in cooperation of the Goethe-University Frankfurt/Main and of the Institute of Lithuanian Language (Vilnius).
[2]The *Referenzcorpus Altdeutsch* is carried through by the Friedrich-Schiller-University of Jena, the Humboldt-University of Berlin, and the Goethe-University of Frankfurt/Main: http://www.deutschdiachrondigital.de/.

Bibliography

Kimbrys, P. (2000). Biblijos vardų žodynas. Aidai, Vilnius.

LKŽ: Lietuvių kalbos žodynas, vol. 1–20. Vilnius, 1941–2002. URL: http://www.lkz.lt/

Palionis, J. (2004). XVI–XVII a. lietuviškų raštų atrankinis žodynas. Mokslo ir enciklopedijų leidybos institutas, Vilnius.

Skardžius, P. (1998[1931]). Die slavischen Lehnwörter im Altlitauischen. In Rinktiniai raštai, vol. 4. Mokslo ir enciklopedijų leidybos institutas, Vilnius. 61–309

Zinkevičius, V. (2000). Lemuoklis–morfologinei analizei. In Darbai ir dienos, vol. 24. Vytauto Didžiojo Universitetas, Kaunas. 245–273. URL: http://donelaitis.vdu.lt/publikacijos/zinkevicius.pdf

Zinkevičius, V., Drożdżyński, W., Homola, P., and Piskorski, J. (2003). Adapting SProUT to processing Baltic and Slavonic languages. In Proceedings of the Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages, held in conjunction with the Conference Recent Advances in Natural Language Processing, 10-12 September 2003, Borovets, Bulgaria. URL: http://www.dfki.de/dfkibib/publications/docs/homola_baltslavir.pdf

Zinkevičius, V. (2008). The Digitization of the Dictionary of the Lithuanian Language. In The Third Baltic Conference on Human Language Technologies (October 4–5, 2007), Kaunas. Vytauto Didžiojo universitetas, Lietuvių kalbos institutas, Vilnius. 349–355

Author Index

Ulrike Demske
Institut für Germanistik,
Universität Potsdam
udemske@uni-potsdam.de

Jolanta Gelumbeckaitė
Institut für Empirische Sprachwissenschaft,
Universität Frankfurt am Main
gelumbeckaite@em.uni-frankfurt.de

Jost Gippert
Institut für Empirische Sprachwissenschaft,
Universität Frankfurt am Main
gippert@em.uni-frankfurt.de

Thomas Jügel
Digital Humanities,
Universität Frankfurt am Main
juegel@stud.uni-frankfurt.de

Sonja Linde
Institut für deutsche Sprache und Linguistik,
Humboldt-Universität Berlin
lindeson@cms.hu-berlin.de

Roland Mittmann
Institut für Empirische Sprachwissenschaft,
Universität Frankfurt am Main
mittmann@em.uni-frankfurt.de

Dennis Pauly
Institut für Germanistik,
Universität Potsdam
denpauly@uni-potsdam.de

Ulyana Senyuk
Institut für Germanistik,
Universität Potsdam
senyuk@uni-potsdam.de

Mindaugas Šinkūnas
Institute of the Lithuanian Language, Vilnius
mindaugas@lki.lt

Vytautas Zinkevičius
Institute of the Lithuanian Language, Vilnius
vytasz@lki.lt