Referenzkorpus Altdeutsch

Kurzbeschreibung

Sonja Linde

www.sprachgeschichte.de/DDD

1 Einleitung

Das **Referenzkorpus Altdeutsch** erfasst und annotiert sämtliche althochdeutschen und altniederdeutschen Texte (ca. 750 – 1050 u.Z.). Nach Abschluss der Arbeiten wird das Korpus etwa 650 000 Wörter umfassen. Das Korpus wird mittels der Datenbank ANNIS[1], die exzellente Recherchemöglichkeiten bietet, der Öffentlichkeit zur Verfügung gestellt.

Die althochdeutsche und altniederdeutsche Überlieferung umfasst 5 größere Texte (Isidor, Tatian, Otfrid, Notker und Heliand) und eine Vielzahl an kleineren und kleinsten Textdenkmälern aus verschiedenen Sprachgebieten. Da ein Teil der ahd. Überlieferung einen engen Bezug zu lateinischen Vorlagen aufweist, werden im Korpus auch die parallel überlieferten lateinische Texte und Textpassagen erfasst, vollständig annotiert und mit den ahd. Daten aligniert, um so Vergleiche zwischen der lat. Vorlage und der volkssprachlichen Übersetzung oder Bearbeitung zu ermöglichen.

Die Grundlage für das Korpus bilden als Referenztexte die handschriftengetreuesten gedruckten Texteditionen, soweit diese zugänglich sind. Alle Wortformen werden sowohl auf der Wortebene als auch auf der Buchstabenebene aufgenommen, so dass es möglich ist, neben den belegten Wortformen auch nach einzelnen Graphemen zu suchen. Neben den Formen der Editionen werden die Schreibweise der Handschriften ebenso wie einheitlich standardisierte Wortformen in das Korpus aufgenommen. Graphische Besonderheiten wie z.B. Rubrizierung in den Manuskripten, Kursivierung, in den Editionen aufgelöste Diakritika oder Ligaturen werden im Korpus kommentiert. Eine Übersicht über die in das Korpus aufgenommenen Referenztexte findet sich im Anhang.

Alle Wortformen sind lemmatisiert. Die Basis für den Lemma-Ansatz des Althochdeutschen bildet das Wörterbuch von Splett 1993; die altsächsische Lemmatisierung orientiert sich an Sehrt 1966. Die Lemmata sind mit neuhochdeutschen Glossierungen versehen, die jeweils an den Kontext angepasst sind.

Die linguistische Annotation umfasst die Wortarten, morphologische Informationen und Angaben zu den Sätzen (siehe 2 – 4), des Weiteren werden Zeilenumbrüche, Absätze und andere Mittel zur Textgliederung ebenso wie Angaben zu Versgliederung und Reimpositionen, soweit vorhanden, erfasst.

[1] ANNIS wurde im Rahmen des SFB 632 Informationsstruktur (Potsdam/Berlin) entwickelt, vgl. http://www.sfb632.uni-potsdam.de/d1/annis/.

Alle in die Datenbank aufgenommenen Texte sind mit Header-Informationen versehen, die die sprach- und literaturwissenschaftlich relevanten Informationen zum jeweiligen Text wie z.B. Entstehungszeit, sprachlicher Raum und Überlieferungskontext enthalten.

2 Tagging der Wortarten

Die Annotation der Wortart wird nach dem von den Partnerprojekten ‚Referenzkorpus Altdeutsch' und ‚Referenzkorpus Mittelhochdeutsch' entwickelten DDDTS (Deutsch Diachron Digital Tagset) vorgenommen. Dieses Tagset schließt sich an das bekannte und gut eingeführte STTS (Stuttgart Tübingen Tagset; Schiller et al. 1999) an und modifiziert dieses für die besonderen Ansprüche der Annotation des historischen Deutschen. Das STTS baut die Zuweisung der einzelnen Wortarten und deren Präzisierungen hierarchisch auf, so dass ein Tag zunächst durch „möglichst selbsterklärende Buchstabensequenzen" (Schiller et al. 1999:4) die Hauptwortart kennzeichnet und sich dann die Markierungen von möglichen Unterwortarten anschließen.

Die Übernahme dieses Vorgehens für DDDTS gewährleistet die klare Kennzeichnung von Hauptwortarten entsprechend STTS, womit die Zusammenführung von STTS-annotierten und DDDTS-annotierten Korpora vor der Perspektive, diachrone Korpora aus den annotiert vorliegenden Materialien zu erstellen, ermöglicht wird. Zum anderen gestattet der Bausatz-artige Aufbau der einzelnen Tags die Modifizierung der Beschreibungen von Unterwortarten, die sich aus den grammatischen Besonderheiten der historischen Texte und unseren Überlegungen ergeben.

Jede Wortart einer Wortform wird zweifach – nach Lemma und nach Beleg – annotiert. Mit diesem Vorgehen wird zwischen der Zuordnung des Lemmas zu einer Hauptwortart und der Verwendung des konkreten Belegs unterschieden. Des Weiteren wird so der oftmals unsicheren Beleglage hinsichtlich der Wortartzugehörigkeit, aber auch möglichen Erscheinungen von Wortklassenwechsel Rechnung getragen.

(1) zeigt die Präzisierung der Wortart des Lemmas an der PoS-Annotation der belegten Wortform. So ist z.B. *in* als Adposition [AP] im Lexikon verzeichnet und wird an dieser Textstelle als Präposition [APPR] realisiert. Die Lemmata der Hilfsverben ‚sein', ‚werden' und ‚haben' werden in DDDTS in Anlehnung an STTS stets als Auxiliare [VA] ge-

kennzeichnet[2]. In (1) wird die finite Form von ‚sein' als Vollverb verwendet, was durch das entsprechende Tagging abgebildet wird.

(1) Tat 88,1

Ist in Hierusalem scáfuuiuuari, the ginemnit ist in ebreiscun Bethsaida
ist in Jerusalem Schafweiher PTK:REL genannt ist in Hebräisch Bethesda
‚Es gibt in Jerusalem einen Schafweiher, der auf Hebräisch Bethesda genannt wird …'

AHD	*Ist*	*in*	*Hierusalem*	*scáfuuiuuari*	*,*	*the*	*ginemnit*	*ist*
Lemma		in	Jerusalem			de		
POS Lemma	VA	AP	NE	NA	$,	PTK	VV	VA
POS Beleg	VVFIN	APPR	NE	NA	$,	PTKREL	VVPP	VAFIN

(…)

Bestimmte Sprachwandelerscheinungen, deren Reflexe sich in den altdeutschen Daten zeigen, können durch die getrennte Annotation der Wortart des Lemmas und der Wortart der belegten Wortform dargestellt werden.

So entwickelt sich z.B. im Althochdeutschen die Verwendung der Genitiv-Plural-Form *iro* des Personalpronomens in der Funktion eines Possessivums, ohne jedoch in das Flexionsparadigma des Possessivpronomens integriert zu sein (2). Synchron ist des Weiteren im Altdeutschen wie in (2) bei – möglichen – Partikelverben oft nicht zweifelsfrei entscheidbar, ob es sich bei den Verbzusätzen um trennbare Verbbestandteile oder um echte Adverben handelt. Bei diesen und weiteren Phänomena wird bei der Kennzeichnung zwischen der Wortart des Lemmas und der des Belegs unterschieden.

(2) Tat 171,3

Uz fon iro samanungu duont sie iuuuih
aus von sie-GEN.PL.M Versammlung-DAT.SG.F tun sie euch
‚Sie schließen euch aus ihrer Gemeinde aus'
(*uz-duon* „heraus-tun" = ausschließen, rauswerfen)

[2] Dieses Vorgehen mag aus diachroner Perspektive problematisch erscheinen, bringt aber den Vorteil mit sich, dass sinnvolle parallele Suchabfragen in DDDTS- und STTS- annotierten Korpora ermöglicht werden.

AHD	*Uz*	*fon*	*iro*	*samanungu*	*duont*	*sie*	*iuuuih*
POS Lemma	ADV	AP	PPER	NA	VV	PPER	PPER
POS Beleg	PTKVZ	APPR	DPOS	NA	VVFIN	PPER	PPER

Deutsch Diachron Digital Tagset (DDDTS)

3 Morphologische Annotation

Die morphologische Annotation kennzeichnet zum einen die unveränderlichen, dem Lemma inhärenten Kategorien wie Flexionsklasse und ggf. Genus, zum anderen werden die veränderlichen flexionsmorphologischen Merkmale erfasst. Die Trennung der veränderlichen von den unveränderlichen Werten lässt für letztere die Unterscheidung von Lemma und der belegten Wortform ähnlich dem Vorgehen beim PoS-Tagging zu, wodurch schwankende Flexionsklassen ebenso wie Genuswechsel erfasst werden können. Bei den adjektivischen Flexionen wird auf diese Weise die Realisierung der starken oder schwachen Flexion an der belegten Wortform vermerkt.

Die Kennzeichnung der morphologischen Werte erfolgt demnach auf drei voneinander unterschiedenen, jeweils einzeln recherchierbaren Ebenen:

1. Flexionsklasse und ggf. Genus des Lemmas
2. Flexionsklasse und ggf. Genus des Belegs
3. Flexionsmorphologie des Belegs

Die folgende Tabelle (3) zeigt die Zuweisung der unveränderlichen Werte nach Wortart jeweils bei Lemma- und bei Belegbezug.

(3) *Tagging der unveränderlichen morphologischen Kategorien*[3]

Attribut	Wortart	PoS-Tag Lemma	mögliche Werte Lemma	mögliche Werte Beleg
Flexionsklasse	Substantiv	NA	*a, ja, wa, z, o, jo, wo, n, in, er, nd, C*	*a, ja, wa, z, o, jo, wo, n, in, er, nd, C*
	Adjektiv	ADJ	*a, ja, wa, o, jo, wo, u*	*a, ja, wa, o, jo, wo, u, P, n*
	Pronomina	DPOS, DW, PI	*a, o*	*a, o, P, n*
	Numerale	CARD	*a, o, i, irr*	*a, o, i, P, n, irr*
		ADJO	*a, o*	*a, o, P, n*
	Verb	VV, VA, VM	*wk1a, wk1b, wk2, wk3, st1a, st1b, st2a, st2b, st3a, st3b, st4, st5, st6, red1, red2, prpr, irr*	*wk1a, wk1b, wk2, wk3, st1a, st1b, st1, st2, st2a, st2b, st3a, st3b, st4, st5, st6, red1, red1a, red1b, red2, prpr, irr*
Genus	Substantiv	NA	*Fem, Masc, Neut*	*Fem, Masc, Neut*

Wenn die Vergabe mehrerer Werte möglich ist, werden alle Möglichkeiten angeführt. Dies ist bei den adjektivischen Flexionen mit lemmatischem Bezug immer der Fall, da diese nach dem femininen (o-) oder nach dem maskulinen (a-) Paradigma flektieren können. Bei Substantiven und Verben werden alle für das jeweilige Lemma im Wortschatz verzeichneten Flexionsklassen annotiert. Für den Bezug auf den Beleg werden nur die an der Wortform flexionsmorphologisch realisierten Flexionsklassen angeführt.

(5) Tat 182, 7

inti mannes sun uuirdit giselit in hant suntigero
und Mannes Sohn wird übergeben in Hand sündig-GEN.PL.M
‚Und der Menschensohn wird in die Hand der Sünder übergeben‘

[3] Die Werte in den Tabellen (3) und (4) beziehen sich ausschließlich auf das deutsche Tagset, d.h. für die lateinischen Daten werden für die unterschiedlichen Kategorien z.T. abweichende und/oder zusätzliche Werte vergeben.

AHD	*Inti*	*mannes*	*sun*	*uuirdit*	*geselit*	*in*	*hant*	*suntigero*
POS Lemma	KO	NA	NA	VA	VV	AP	NA	ADJ
POS Beleg	KON	NA	NA	VAFIN	VVPP	APPR	NA	ADJS
Flexion Lemma		C,a_Masc	u,i_Masc	st3b	wk1a		u,i_Fem	a,o
Flexion Beleg 1		a_Masc	i_Masc	st3b	wk1a		i_Fem	P

Tabelle (6) gibt die flexionsmorphologischen Werte, die für eine belegte Wortform vergeben werden können, an.[4]

(6) *Tagging der veränderlichen morphologischen Kategorien*

Attribut	Wortart	PoS-Tag Beleg	mögliche Werte
Numerus	Substantiv, Adjektiv, Pronomen, Verb, Numerale	NA, NE, NEO; ADJ, ADJN, ADJD, ADJS, ADJO; PI, PIN; DI, DIA, DIN, DIS; DPOS, DPOSN, DPOSS; DW, DWS, DWSREL; VVFIN, VVINFS, VVPS, VVPSS, VVPSD, VVPSA, VVPSN, VVPP, VVPPS, VVPPD, VVPPA, VVPPN, VVIMP, VAFIN, VAIMP, VMFIN; CARD, CARDN, CARDS	*Sg, Pl, Du*
Kasus	Substantiv, Adjektiv, Pronomen, Numerale, Verb	NA, NE, NEO; ADJ, ADJN, ADJD, ADJS, ADJO; PI, PIN; DI, DIA, DIN, DIS; DPOS, DPOSN, DPOSS; DW, DWS, DWSREL; CARD, CARDN, CARDS; VVINFS	*Nom, Gen, Dat, Acc, Ins*

[4] Hierbei bestehen zwischen altsächsischen und den althochdeutschen Werten z.T. Unterschiede, so dass nicht alle Werte für beide Sprachen verwendet werden.

Grad	Adjektiv, Verb	ADJN, ADJD, ADJS, ADJO; VVPS, VVPSS, VVPSD, VVPSA, VVPSN, VVPP, VVPPS, VVPPD, VVPPA, VVPPN	*Pos, Comp, Sup*
Deklinations-weise	Adjektiv, Pronomen	ADJN, ADJD, ADJS, ADJO; DPOS, DPOSS, DPOSN	*st, sw, 0*
Modus	Verb	VVFIN, VVIMP, VAFIN, VAIMP, VMFIN	*Ind, Subj, Imp*
Tempus	Verb	VVFIN, VAFIN, VMFIN	*Pres, Past*
Person	Pronomen, Verb	DPOS, DPOSS, DPOSN; VVFIN, VAFIN, VMFIN	*1, 2, 3*
Partizipal-flexionsklasse	Verb	VVPS, VVPSS, VVPSD, VVPSA, VVPSN, VVPP, VVPPS, VVPPD, VVPPA, VVPPN	*ja, jo, a, o, P, n*

Die Vergabe der veränderlichen flexionsmorphologischen Werte erfolgt wie beschrieben auf einer zusätzlichen Ebene.

(5‘) Tat 182, 7

(…)

AHD	*mannes*	*sun*	*uuirdit*	*geselit*	*in*	*hant*	*suntigero*
POS Beleg	NA	NA	VAFIN	VVPP	AP	NA	ADJS
Flexion Beleg 2	Sg_Gen	Sg_Nom	Ind_Pres_Sg_3			Sg_Acc	Pos_Masc_Pl_Gen_st

Alle drei morphologischen Ebenen können separat durchsucht werden. Es besteht also z.B. die Möglichkeit, über einfache statistische Abfragen zu erfahren, wie häufig ein bestimmtes, in verschiedenen Klassen belegtes ahd. Lemma in den overten Kasus welcher Klasse folgt.

4 Annotation der Sätze

Für das Referenzkorpus Altdeutsch ist in der ersten Phase keine syntaktische Kommentierung mittels einer Baumbank-Annotation vorgesehen, jedoch werden bestimmte syntaktische Informationen mithilfe eines einfachen, linearen Schemas angeboten, um so z.B. die Suche nach bestimmten Satztypen zu ermöglichen.

Jeder Satz wird als Spanne gekennzeichnet, welche mit Werten für z.B. Finitheit, syntaktischer Status oder Semantik des Satzes versehen ist.

Die Satzannotation erfolgt mit den in (7) aufgeführten Werten:

(7)

Tag	Beschreibung
Ad	*adversative clause*, Adversativsatz
Adv	*adverbial clause*, Adverbialsatz
Att	*attributive clause*, Attributsatz
Caus	*causal clause*, Kausalsatz
CC	*continued clause*, Wiederaufnahme
CE	*elliptical clause*, elliptischer Satz
CF	*finite clause*, finiter Satz
CI	*infinite clause*, infiniter Satz
Cnc	*concessive clause*, Konzessivsatz
Cnd	*conditional clause*, Konditionalsatz
Cns	*consecutive clause*, Folgesatz
CP	*participle clause*, Partizipalsatz
CS	*suspended clause*, Unterbrechung
Ex	*exclamation*, Ausrufesatz
Fin	*final clause*, Finalsatz
I	*introduction*, eingeleiteter Satz
ID	*dependent interrogative clause*, abhängiger Fragesatz
Imp	*imperative clause*, Imperativsatz
Int	*interrogative clause*, Fragesatz
Loc	*local clause*, Lokalsatz
M	*main clause*, Hauptsatz
Mod	*modal clause*, Modalsatz
O	*object clause*, Objektsatz

P	*predicative clause*, Prädikativsatz
S	*subject clause*, Subjektsatz
Temp	*temporal clause*, Temporalsatz
U	*clause unintroduced*, uneingeleiteter Satz

Die Werte über die syntaktischen Informationen werden stets in der Reihenfolge wie in (8) vergeben:

(8)

CF	CS	I	M	Temp
CI	CC	U	S	Loc
CP			O	Caus
CE			Att	Mod
			P	Fin
			Int	Cnd
			ID	Cnc
			Imp	Cns
			Ex	Ad
			Adv	

Ein eingeleiteter Hauptsatz wie in (5) wird also wie folgt kommentiert:

(9) *inti mannes sun uuirdit giselit in hant suntigero*

CF_I_M

(10) zeigt die Annotation eines komplexen Satzgefüges mit allen in der Datenbank durchsuchbaren Ebenen.

(10) Tat 168, 3

Ih	*quidu*	*iuuuih*	*friunta*	*uuanta*	*allu*	*thiu*	*ih*	*gihorta*
ich	sage	euch-ACC	Freunde-ACC	weil	alle-ACC.N	die-ACC.PL.N	ich	hörte

fon	*minemo*	*fater*	*teta*	*ih*	*iu*	*cundiu*
von	meinem	Vater	tat	ich	euch-DAT	kund-ACC.PL.N

‚Ich sage von euch, dass ihr Freunde seid, weil ich Euch alles, was ich von meinem Vater gehört habe, bekannt mache‘

AHD	*ih*	*quidu*	*iuuuih*	*friunta*	,	*uuanta*	*allu*	,
Lemma	ich	quëdan	ir	friunt		wanta	al	
POS Lemma	PPER	VV	PPER	NA	$,	KO	DI	$,
POS Beleg	PPER	VVFIN	PPER	NA	$,	KOUS	DIS	$,
Flexion Lemma		st5		a,nd_Masc			a,o	
Flexion Beleg		st5		a_Masc			P	
Flexion Beleg 2	Sg_Nom_1	Ind_Pres_Sg_1	Pl_Acc_2	Pl_Acc			Neut_Pl_Acc_st	
Satz	CF_U_M					CF_CS_I_Adv_Caus		

AHD	*thiu*	*ih*	*gihorta*	*fon*	*minemo*	*fater*
Lemma	dër	ich	gihōren	fona	mīn	fater
POS Lemma	DD	PPER	VV	AP	DPOS	NA
POS Beleg	DDSREL	PPER	VVFIN	APPR	DPOS	NA
Flexion Lemma			wk1a		a,o	er_Masc
Flexion Beleg 1			wk1a		P	er_Masc
Flexion Beleg 2	Neut_Pl_Acc	Sg_Nom_1	Ind_Past_Sg_1		Masc_Sg_Dat_st	Sg_Dat
Satz	CF_I_Att					

AHD	*teta*	*ih*	*iu*	*cundiu*	.
Lemma	tuon	ih	ir	kund	
POS Lemma	VV	PPER	PPER	ADJ	$.
POS Beleg	VVFIN	PPER	PPER	ADJD	$.
Flexion Lemma	irr			a,o	
Flexion Beleg 1	irr			P	
Flexion Beleg 2	Ind_Past_Sg_1	Sg_Nom_1	Pl_Dat_2	Pos_Neut_Pl_Acc_st	
Satz	CF_CC_I_Adv_Caus				

5 Literatur

Schiller, Anne et al. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS. (Großes und kleines Tagset)*. http://www.sfb441.uni-tuebingen.de/a5/codii/info-stts-en.xhtml

Sehrt, Edward H. (1966). *Vollständiges Wörterbuch zum Heliand und zur altsächsischen Genesis*. Göttingen: Vandenhoeck und Ruprecht.

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

Anhang: Referenztexte