Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

REFERENZKORPUS MITTELNIEDERDEUTSCH/ NIEDERRHEINISCH 1200 - 1650

Grenzfälle der Annotation historischer Daten

Katharina Dreessen / Sarah Ihden/ Timm Lehmberg

Clarin-D-Workshop: Annotation

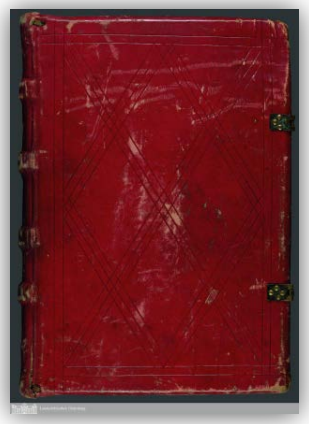
29. November 2013

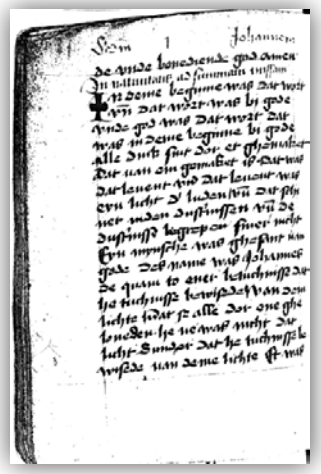
Design des Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200-1650)

Schreibsprach-landschaften	Zeiträume	Felder der Schriftlichkeit
Nordniedersächsisch	I: 1200-1300	Verwaltung
Ostelbisch	II: 1301-1350	Recht
Baltisch	III: 1351-1400	Urkunden
Westfälisch	IV: 1401-1450	Wissensvermittlung
Ostfälisch	V: 1451-1500	Religion (Kirchliches)
Elbostfälisch	VI: 1501-1550	Literarische Texte
Südmärkisch	VII: 1551-1600	Alltagsschriftlichkeit
Niederrheinisch	VIII: 1601-1650	Inschriften

Korpuserstellung und -aufbereitung: Teil I

Originaltext > Reproduktion > Transkription





$Kap.1$
$Bl.1$
S{e}c{un}d{u}m j Johannem
Jn natiuitatis ad summam missam
$V.1$ In deme beginne was dat wort $.$
vn{de} dat wort was bi gode $.$
vnde god was dat wort $.$ $V.2$ dat
was in deme beginne bi gode $.$
$V.3$ alle dink sint dor et ghemaket $.$
Dat uan em gemaket is $.$ $V.4$ Dat was
dat leuent $.$ vn{de} dat leuent was
eyn licht d{er} luden / $.$ $V.5$ vn{de} dat schi#
net in§den dust{er}nissen $.$ vn{de} de
dust{er}nisse begrepen siner nicht $.$

Korpuserstellung und -aufbereitung: Teil II

Lemmatisierung > Annotation > Publikation

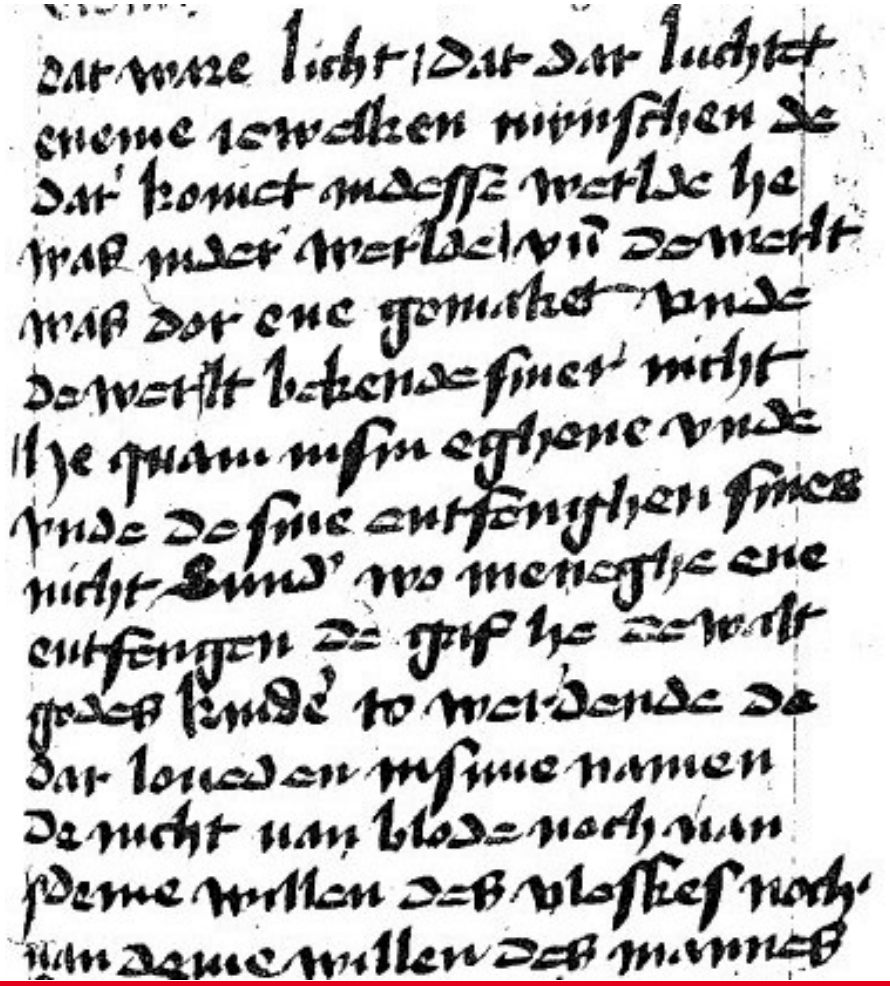
ID	Lexem	Lemma
11	$V.1$	ED
12	In	in
13	deme	dat
14	beginne	begin
15	was	sîn: wesen
16	dat	dat
17	wort	wort
18	$.$	ED
19	vn{de}	unde
20	dat	dat
21	wort	wort
22	was	sîn: wesen
23	bi	bî
24	gode	god
25	$.$	ED
26	vnde	unde
27	god	god
28	was	sîn: wesen
29	dat	dat
30	wort	wort
31	$.$	ED

ID	Lexem	Wortart
11	$V.1$	ED
12	In	APPR
13	deme	DAD
14	beginne	NA
15	was	VVFIN
16	dat	DAD
17	wort	NA
18	$.$	ED
19	vn{de}	KON
20	dat	DAD
21	wort	NA
22	was	VVFIN
23	bi	APPR
24	gode	NA
25	$.$	ED
26	vnde	KON
27	god	NA
28	was	VVFIN
29	dat	DAD
30	wort	NA
31	$.$	ED

Automatische Annotation (POS-Tagging): Workflow im Überblick

Automatisches Part-Of-Speech (POS) Tagging als iterativer Prozess

Vor der Annotation: Präeditierung

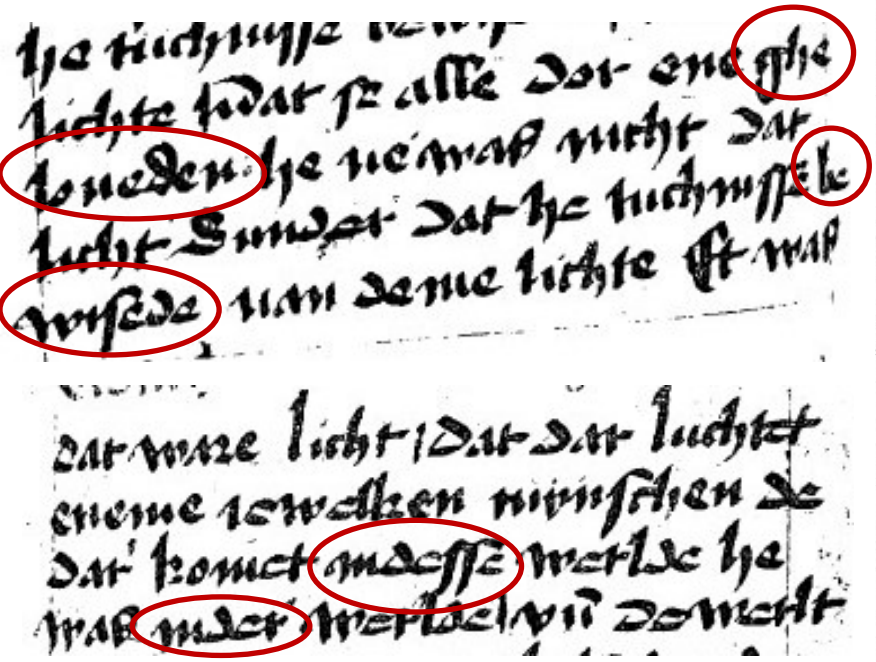


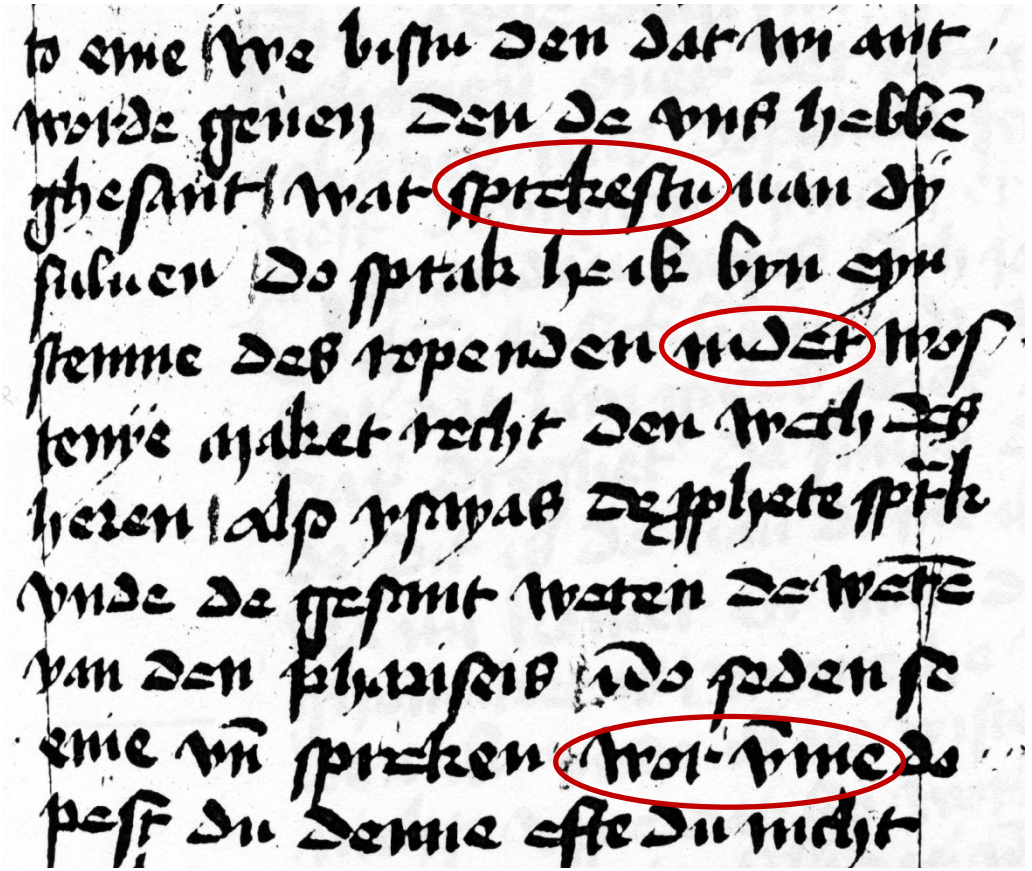
Satzgrenzenfestlegung

- Vor automatischer Annotation: Satzgrenzenfestlegung
- Die Bestimmung einer Satzeinheit richtet sich nach dem Vorhandensein eines finiten Verbes.
- Das Ende einer jeden Satzeinheit wird mit $.$ gekennzeichnet.
- Durch eingeschobene Sätze getrennte Satzeinheiten werden durch $.a$ und $.a$ gesondert ausgewiesen, um sie in späteren Abfragen einander zuordnen zu können.

$V.15$ Johannes bewisede tuchnisse uam
em $.$ vn{de} ropet sprekende $.$ Dit
is de $.$ den ik sprak $.$ de na my kome{n}=
de is $.$ de is uor mi ghemaket $.$ we{n}=
te he was er den ik / $.$ $V.16$ vn{de} uan sin{er}
vullenkomenheit hebbe wi alle
nomen gnade v{m}me gnade / $.$ $V.17$ we{n}te
de . e . is ghegheuen dor moyses $.$
gnade vn{de} warheit is ghemaket
dor ih{esu}m {christum} / $.$ $V.18$ god ne heuet ne#
man ghesien $.$ de enboren sone $.a$
de dar is in§deme schatte des ua=
der $.$ de seget et $.b$ *Domi{ni}ca iiij post*

Zusammen- und Getrenntschreibung





Zusammen- und Getrenntschreibung – Beispiele

Getrenntschreibung eines zusammengehörenden Lexems

Komposita: denest lude

Präfixe: ghe heten

be kende

Zusammenschreibung zweier (oder mehrerer) Lexeme

Schreiberbedingte (Einzel-)Fälle: lamgodes

Klitisierung: enmohte

inder

Krasis: sprekestu

Worttrennung am Zeilenende

nicht oder nicht einheitlich und konsequent gekennzeichnet

Zusammen- und Getrenntschreibung

Vor Segmentierung: Normalisierung im Bereich der Zusammen- und Getrenntschreibung

lichte $.$ Dat se alle dor ene ghe#
loueden $.$ $V.8$ he ne was nicht dat
licht $.$ Sunder dat he tuchnisse be#
wisede uan deme lichte $.$ $V.9$ Et was
$Bl.2$
S{e}c{un}d{u}m ii Joha{n}nem
Joh{annes} I .
dat ware licht / $.$ dat dar luchtet
eneme iewelken mynschen $.$ de
dar komet in§desse werlde $.$ $V.10$ he
was in§der werlde / $.$ vn{de} de werlt

to eme / $.$ we bistu den $.$ dat wi ant=
worde geuen den $.$ de vns hebbe{n}
ghesant / $.$ wat sprekestu uan dy
suluen / $.$ $V.23$ do sprak he $.$ ik byn eyn
stemne des ropenden in§der wos=
tenye $.$ maket recht den wech des
heren / $.$ also ysayas de {pro}phete spr{a}k $.$
$V.24$ vnde de gesant weren $.$ de were{n}
van den phariseis / $.$ $V.25$ Do seden se
eme $.$ vn{de} spreken $.$ wor# v{m}me do=
pest du denne $.$ efte du nicht

Grenzfälle automatischer Annotation: 1) zusammengeschriebene Worteinheiten

	A	B	C
1	**ID**	**Lexem**	**Wortart**
2	112	Dat	KOUS
3	113	se	PPER
4	114	alle	PI
5	115	dor	APPR
6	116	ene	PPER
7	117	ghe#loueden	VVFIN
8	118	$.$	ED
9	119	$V.8$	ED
10	120	he	PPER
11	121	ne	PTKNEG
12	122	was	VVFIN
13	123	nicht	PTKNEG
14	124	dat	DAD
15	125	licht	NA
16	126	$.$	ED
17	127	Sunder	KON
18	128	dat	KOUS
19	129	he	PPER
20	130	tuchnisse	NA
21	131	be#wisede	VVFIN
22	132	uan	APPR
23	133	deme	DAD
24	134	lichte	NA
25	135	$.$	ED

	A	B	C
1	**ID**	**Lexem**	**Wortart**
2	462	Do	ADV
3	463	sprak	VVFIN
4	464	he	PPER
5	465	$.$	ED
6	466	ik§ne	PPER;PTKNEG
7	467	bin	VVFIN
8	468	/	$
9	469	$.$	ED
10	470	Do	ADV
11	471	spreken	VVFIN
12	472	se	PPER
13	473	$.$	ED
14	474	bistu	VVFIN;PPER
15	475	en	DAI
16	476	{pro}phete	NA
17	477	$.$	ED

Grenzfälle automatischer Annotation: 2) getrennte Worteinheiten

	A	B	C
1	**ID**	**Lexem**	**Wortart**
2	4592	sta	VVFIN
3	4593	vp	PTKVZ
4	4594	$.$	ED
5	4595	vn{de}	KON
6	4596	bore	VVFIN
7	4597	vp	PTKVZ
8	4598	din	DPOS
9	4599	bedde	NA
10	4600	$.$	ED
11			
12	14679	alle	DI
13	14680	dat	DAD
14	14681	volk	NA
15	14682	is	VAFIN
16	14683	na	APPR
17	14684	em	PPER
18	14685	gan	VVPP
19	14686	$.$	ED
20	14687	$V.20$	ED
21	14688	su{n}=der	KON
22	14689	dar	ADVPROL
23	14690	weren	VVFIN
24	14691	etesweke	PI
25	14692	heydene{n}	NA
26	14693	van	ADVPROR
27	14694	$.$	ED

	A	B	C	D
1	**ID**	**Lexem**	**Wortart**	**Kommentar**
2	1940	Et	PPER	
3	1941	was	VVFIN	
4	1942	en	DAI	
5	1943	mynsche	NA	
6	1944	ua{n}	APPR	
7	1945	den	DAD	
8	1946	pha#	NA	getrenntes Lexem A
9	1947	*C{apitel}	NA	
10	1948	III	XY	
11	1949	.*	$	
12	1950	riseis	NA	getrenntes Lexem B
13	1951	$.$	ED	
14				

Grenzfälle automatischer Annotation: 2) sprachspezifisches Tagging

- Herausforderung: mangelnde Datengrundlage für das Mittelniederdeutsche

  -> je umfassender die Datengrundlage, desto zuverlässiger ist das Ergebnis des Taggers und desto weniger Kontrolldurchgänge sind nötig

  -> Tagger beginnt erst anhand der ins Korpus aufgenommenen Texte zu lernen
- Intention: 1) Tagset ist differenziert genug, um alle notwendigen Fälle zu erfassen und den Aufwand potentieller Nachannotationen gering zu halten

  2) Tagset ist nicht zu differenziert, um so eine möglichst hohe Erfolgsquote des automatischen Taggings zu gewährleisten und so den Kontrollaufwand gering zu halten
- Ergebnis: für ReN spezifisches Tagset (auf STTS basierend, stark an HiTS orientiert)

Tagset für ReN

	POS-Kategorien	Beschreibung
Adjektive	ADJA	attributives Adjektiv
	ADJV	adverbiales Adjektiv
	ADJP	prädikatives Adjektiv
	ADJO	Ordinalzahlen
Adverbien	ADV	lokale Adverbien temporale Adverbien modale Adverbien kausale Adverbien Ordinalzahlen Multiplikativzahlen unbestimmte Gattungszahlen
	ADVNEG	Adverbien, negativ
	ADVREL	Relativadverbien
	ADVW	Interrogativadverbien
	ADVKOM	Kommentaradverbien
	ADVPRO	Pronominaladverbien
	ADVPROL	Pronominaladverbien, getrennt, linker Teil
	ADVPROR	Pronominaladverbien, getrennt, rechter Teil
	ADVKON	Konjunktionaladverbien
Adpositionen	APPR	Präposition, präpositionsähnliche Adjektive
	APPRART	Präposition mit inkorporiertem Artikel
	APPO	Postposition
	APZL	Zirkumposition, linker Teil
	APZR	Zirkumposition, rechter Teil
Zahlen	CARD	Kardinalzahlen
Determinativa	DAD	Definitartikel
	DAI	Indefinitartikel
	DD	demonstratives Artikelwort
	DPOS	possessives Artikelwort
	DI	indefinites Artikelwort (mit oder ohne Determinativ vorkommend)
	DINEG	indefinites Artikelwort, negativ
	DREL	relatives Artikelwort
	DW	interrogatives Artikelwort
Fremdsprachliches Material	FM	fremdsprachliches Material
Interjektionen	ITJ	Interjektion

Konjunktionen	KON	nebenordnende Konjunktion
	KOUI	unterordnende Konjunktion mit Infinitiv
	KOUS	unterordnende Konjunktion mit Satz
	KOKOM	Vergleichspartikel
Nomen	NA	Appellativa
	NE	Eigennamen
Partikeln	PTKZU	Partikel „zu“
	PTKNEG	Negationspartikel
	PTKVZ	abgetrennter Verbzusatz
	PTKA	Partikel bei Adjektiv oder Adverb
	PTKGRD	Gradpartikel
	PTKFOK	Fokuspartikel
	PTKMOD	Modalpartikel (Abtönungspartikel)
	PTKGSP	Gesprächspartikel
Pronomen	PPER	Personalpronomen
	PRF	reflexives Personalpronomen
	PD	Demonstrativpronomen
	PPOS	Possessivpronomen
	PI	Indefinitpronomen
	PINEG	Indefinitpronomen, negativ
	PREL	Relativpronomen
	PW	Interrogativpronomen
Verben	VAFIN	finites Hilfsverb
	VMFIN	finites Modalverb
	VMINF	Modalverb im Infinitiv (Ersatzinfinitiv)
	VVFIN	finites Vollverb
	VVINF	Vollverb im Infinitiv
	VVIMP	Vollverb im Imperativ
	VVPP	Vollverb im Part. Prät.
	VVPS	Vollverb im Part. Präs.
	VVIZU	Vollverb im Infinitiv mit eingeschobenem „zu“ (Partikelverb)
Interpunktion	$	Interpunktion
Nichtwörter	XY	Nichtwort
Ohne Annotat	OA	nicht annotierbare Lexeme, z.B. aufgrund von Schreib- oder Druckfehlern
Editorische Angaben	ED	Editorische Angaben

Grenzfälle automatischer Annotation: 3) formales und/oder funktionales POS-Tagging/ Wortartwechsel

Form vs. Funktion

Beispiel: Vn{de} se crucegede{n} myt eme twe morde{re} $.$

Dar se he{n}ge{n} […] ihesum myddes $.$

formal: ADVPRO funktional: ADVREL

-> Lösung: ADVREL<ADVPRO

Wortartwechsel

Beispiele: 1) de ogen des blinden

NA<ADJ

2) de walt godes kindere to werdende

NA<VVINF

3) eyn born des sprengenden waters

ADJA<VVPS

Vielen Dank für Ihre Aufmerksamkeit!