Bochumer
Linguistische
Arbeitsberichte
19

Handbuch zum Referenzkorpus Mittelhochdeutsch

Thomas Klein und Stefanie Dipper

Bochumer Linguistische Arbeitsberichte

Herausgeberin: Stefanie Dipper

Die online publizierte Reihe „Bochumer Linguistische Arbeitsberichte“ (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series “Bochumer Linguistische Arbeitsberichte” (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

Band 19 (December 2016)

Herausgeberin: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Erscheinungsjahr 2016
ISSN **2190-0949**

Thomas Klein und Stefanie Dipper

Handbuch zum Referenzkorpus Mittelhochdeutsch

2016

Bochumer Linguistische Arbeitsberichte

(BLA 19)

Inhalt

Dieses Handbuch beruht großenteils auf dem Bonner Korpushandbuch (Projekt „Mhd. Grammatik“), das im Wesentlichen von Tobias Kemper verfasst und stellenweise durch Stefan Müller, Anja Miklin, Pia-Ramona Wojtinnek und Thomas Klein ergänzt wurde. Zu den Abschnitten zur Darstellung in ANNIS haben Adam Roussel, Florian Petran und Marcel Bollmann beigetragen.

1. Lizenz und Zitierweise

Wenn Sie das Korpus zitieren möchten, bitte das folgende Format benutzen:

Klein, Thomas; Wegera, Klaus-Peter; Dipper, Stefanie; Wich-Reif, Claudia (2016). Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0, https://www.linguistics.ruhr-uni-bochum.de/rem/. ISLRN 332-536-136-099-5.

2. Entstehung und Zusammensetzung des ReM-Korpus

Im Referenzkorpus Mittelhochdeutsch (ReM im weiteren Sinne) sind vier verschiedene Korpora zusammengeflossen:

2.1 Das Kölner Korpus hessisch-thüringischer Texte

Von 1986 bis 1993 wurde am Institut für deutsche Sprache und Literatur in Köln unter Leitung von Thomas Klein und Joachim Bumke eine Gruppe von hessisch-thüringischen Texten digitalisiert (M005, M008, M100, M103, M107S, M158, M213, M301, M541H, M541B, M541S) und mit einem von Thomas Klein entwickelten Programmpaket halbautomatisch annotiert (Klein, 1991; 2001). Mit verbesserten Versionen dieses Scriptpakets sind auch alle anderen Teilkopora von ReM annotiert worden. In gedruckter Form wurde das hessisch-thüringische Textkorpus 1997 publiziert (Klein & Bumke, 1997).

Alle diese in Köln annotierten 11 Texte (176.289 Tokens) sind teils in MiGraKo, teils in ReM (im engeren Sinne) übernommen worden.

2.2 Das Bonner Korpus mitteldeutscher Texte

Ab 1993 setzte Thomas Klein den bereits in Köln begonnenen Aufbau einer weiteren annotierten mitteldeutschen Textgruppe zusammen mit Mitarbeitern des (ehem.) Instituts für geschichtliche Landeskunde der Rheinlande in Bonn fort. Schwerpunkte waren zum einen mitteldeutsch-niederdeutsche Texte (M046, M199A, M199B, M199C, M205A, M205S, M206, M240B), zum andern mittelfränkische und weitere westmitteldeutsche Texte (M006, M007, M013O, M013B, M045, M070, M073, M082, M108M, M108P, M108S, M114, M117, M118, M135, M136, M148, M239, M243, M251, M303, M327, M335, M354, M505, M507, M543, M546, M547); die mittelfränkischen Urkunden von M544 wurden großenteils, die moselfränkischen Urkunden von M545 komplett von Britta Weimann annotiert (Weimann, 2012).

Alle diese in Bonn annotierten 39 Texte (260.470 Tokens) sind teils in MiGraKo, teils in eReM übernommen worden.

2.3 Das Bochumer Mittelhochdeutschkorpus (BoMiKo) und das Korpus der Mittelhochdeutschen Grammatik (MiGraKo)

Für das Ziel einer neuen korpusbasierten mittelhochdeutschen Grammatik entwarf Klaus-Peter Wegera Anfang der 1990er Jahre das Design für ein nach Zeit, Sprachraum und Textart strukturiertes Korpus mit strengen Bedingungen für die Textauswahl (Wegera, 1991; 2000). Nach diesen Prinzipien wurde unter Leitung von Klaus-Peter Wegera in Bochum das Bochumer Mittelhochdeutsch-Korpus“ (BoMiKo) zusammengestellt. Dieses Textkorpus bildete die Ausgangsbasis für die DFG-Projekte „Korpus einer mittelhochdeutschen Grammatik“ (1997–1999) und „Mittelhochdeutsche Grammatik“ in den Arbeitsstellen Bochum (Leitung: Klaus-Peter Wegera), Bonn (Leitung: Thomas Klein) und Halle (Leitung: Hans-Joachim Solms). Die Korpusstruktur und Textauswahl wurde im Rahmen dieser Projekte modifiziert und erweitert, so insbesondere durch die Ermittlung geeigneter Urkundenstrecken in der Bochumer Arbeitsstelle (M344–353). Das so entstandene Textkorpus wurde 1997–1999 in den drei Projektarbeitstellen Bochum, Bonn und Halle transkribiert und anschließend in Bonn komplett tokenisiert, interpungiert, lemmatisiert, grammatisch annotiert und graphophonemisch analysiert. Außerdem wurden hier die aus dem Kölner und Bonner Mitteldeutsch-Korpus übernommenen annotierten Texte integriert. Das Ergebnis ist das Korpus der Mhd. Grammatik (MiGraKo). Es umfasst 102 Texte mit ca. 1 Mio. Wortformen.

Strukturiert ist das MiGraKo im Gefolge des BoMiKo zeitlich in Abschnitte von 50 Jahren; nur die quellenarme Frühzeit von 1050–1150 ist zu einem Zeitschnitt zusammgefasst. Sprachräumlich ist das Korpus in 7 (9) Sprachräume und nach Textart in Vers-, Prosa- und Urkundentexte gegliedert:

Vers	Prosa	Urkunde

	westmitteldeutsch			hessisch-thüringisch		
²12	ArnM, RBib	TrPs	–	Aegi	PrFr	–
¹13	RhMl, RhTun	VatG	–	GRud, AlxS	PrMK	–
²13				AthP		

	mittelfränkisch			rheinfränkisch-hessisch			ostmitteldeutsch		
²13	Lilie, KuG	Lilie, Brig	Köln	Himlf,PrRei	SalH, PrM	–	–	MüRB, JMar	
¹14	Yol, Göll	Taul, BuMi	Köln	Elis, Erlös	OxBR, Hleb	Mainz	LuKr, HTri, Pass	BeEv, MBeh	Jena

	ostfränkisch		
¹14	Renn, Lupo	GnaÜ, SBNü, WüPo	Nürnb

		alemannisch			ostalemannisch-westbairisch (alem.-bair. Übergangsraum)			bairisch		
²11/¹12	obd.	Ezzo/Mem, Meri, RPaul				Will, WNot, BaGB/HuH				
²12		LEntc, Scop	PrZü, Muri	–	Mess	Spec, WMEv	–	Kchr, HLit	Phys, Wind	–
¹13	obd.	Iw, Nib, Parz, Tris								
¹13		Flor, TriF	TrHL, Luci	–	–	ZwBR, Hoff	–	Mar, Hchz	PrMi, PrPa	–
²13		RWchr, RWh	SwSp, PrSch	Freib	Wins	DvATr, StBA	Augsb	Diet, Lieht	BKön, Bart	–
¹14		Rapp, Mart	NikP	Freib	Türh	Baum, Hartw	Augsb	MMag	Rupr, ObEv	Lands

2.4 Das Referenzkorpus Mittelhochdeutsch (eReM)

Das „Referenzkorpus Mittelhochdeutsch (1050–1350)“ im engeren Sinne (eReM) entstand 2009–2014 im Rahmen des gleichnamigen DFG-Projekts in Bonn (Leitung Thomas Klein, Claudia Wich-Reif) und Bochum (Leitung Klaus-Peter Wegera, Stefanie Dipper). Das eReM-Korpus wurde von Thomas Klein so konzipiert, dass es MiGraKo in zweifacher Hinsicht ergänzt:

(1) Einerseits wurden möglichst alle frühmhd. Texte aufgenommen, sofern sie nicht schon im MiGraKo enthalten waren. Den strengen für MiGraKo geltenden Aufnahmebedingungen genügen sehr viele dieser Texte nicht, sei es, weil sie zu geringen Umfangs oder schreibsprachlich heterogen sind, sei es, weil sie nicht in einer hinreichend zeitnahen handschriftlichen Überlieferung vorliegen.
Nicht aufgenommen sind bisher nur einige wenige frühmhd. Texte, insbesondere solche, die lediglich fragmentarisch mit vielen Textlücken überliefert sind (so z.B. ‚Trost in Verzweiflung‘; ‚Cantilena de conversione Sti. Pauli‘; ‚Klagenfurter Gebete‘, ‚Schlägler Predigtbruchstücke‘).
In aller Regel sind die Texte komplett erfasst und annotiert worden. Bei folgenden Großtexten wurden jedoch nur zusätzliche Ausschnitte zu den bereits in MiGraKo enthaltenen Ausschnitten des Textes bearbeitet und so der annotierte Teil des Textes entweder komplettiert oder auf wenigstens 20.000 Tokens vergrößert: M113y (zu MiGraKo M113), M188y (zu MiGraKo M188), M195y (zu MiGraKo M195), M214y (zu MiGraKo M214), M241y (zu MiGraKo M241), M242y (zu MiGraKo M242).

(2) Andererseits sollten aus dem Zeitraum von 1200–1350 zur Ergänzung des MiGraKo Texte von ähnlicher hoher Zuordnungsqualität wie der der MiGraKo-Texte selbst aufgenommen werden. Ausnahmen hinsichtlich der Zuordnungsqualität gibt es vor allem bei den in eReM aufgenommenen Texten aus den Köln–Bonner Mitteldeutsch-Korpora, so vor allem M541H, die erst 1333 (also mehr als 100 Jahre nach der Textentstehung) im mainfränkischen Würzburg geschriebene Hs. von Herborts Trojaroman. Da M541H aber in einer konservativen Schreibsprache gehalten ist, die jener des mitteldeutschen Fragments M541B (13. Jh.) noch recht nahesteht, erscheint diese Ausnahme gerechtfertigt.

Außerdem sind wie im frühmhd. Bereich (1) einige der MiGraKo-Prosatexte des 13. Und 14. Jh. in eReM um zusätzliche Ausschnitte erweitert worden: M161 (zu MiGraKo M329), M402Y (zu MiGraKo M402), M403Y (zu MiGraKo M403), M405Y (zu MiGraKo M405), M406Y (zu MiGraKo M406), M407Y (zu MiGraKo M407).

Aus diesen und den unter (1) genannten erweiterten Prosatexten lässt sich ein Korpus von Prosatexten mittleren Umfangs zusammenstellen, das vor allem für Syntaxuntersuchungen geeignet sein dürfte:

Tokens	Siglen (Tokens)	Texttitel	Dialekt – Zeit
21891	M242 (17404) + M242Y (4487)	Wiener Notker	bair - 11,E-12,A
29742	M214 (14422) + M214Y (15320)	Speculum ecclesiae B	mittelbair - um/nach 1200

23488	M113 (10998) + M113Y (12490)	St. Trudperter Hohelied	alem - 13,1V
17718	M329 (13727) + M161 (3991)	Millstätter Predigtsamm-lung	bair - 13,M
19251	M403 (12333) + M403Y (6918)	Buch der Könige	bair - 13,4V (80er Jahre)
18675	M405 (13105) + M405Y (5570)	David von Augsburg: Traktate	mittelbair (westlich) - kurz vor/um 1300
25964	M407 (14639) + M407Y (11325)	Hermann von Fritzlar: Heiligenleben	hess-thür - 1349?
23842	M402 (16457) + M402Y (7385)	Berliner Evangelistar	thür-obersächs - 14,M
18130	M406 (12354) + M406Y (5776)	Christine Ebner: Von der Gnaden Überlast	ofrk, nordbair - um/kurz nach 1350

eReM ist also für sich genommen kein strukturiertes Korpus wie MiGraKo. Es bietet zum einen aber die Möglichkeit, MiGraKo durch Hinzunahme strukturell passender e-ReM-Texte oder -Textausschnitte zu erweitern, und zum andern die Möglichkeit nahezu flächendeckender Recherchen im Bereich des Frühmittelhochdeutschen.

2.5 Bezeichnung der Textgruppen von eReM und MiGraKo in ANNIS

In ANNIS erscheinen die MiGraKo- und eReM-Texte in Textgruppen, die sich weitmöglichst an der MiGraKo/BoMiKo-Struktur orientieren (s. 2.3). Die Namen der Textgruppen sind folgendermaßen zu lesen:

12_2	-	**bair**	-	**PV**	-	**G** (oder **X**)
Zeit-raum		Sprach-raum		Textart		Korpus

- Die Buchstaben G und X bezeichnen die beiden Korpora G = MiGraKo (Grammatikkorpus) und X (eXtensions) = eReM.
- Es folgt die Bezeichnung des Zeitabschnitts, z.B. „13_2“ = 13. Jh., 2. Hälfte; „11-12_1“ = 2. Hälfte 11. Jh. bis 1. Hälfte 12. Jh.
- Durch „-“ abgetrennt schließt sich der Sprachraum in üblicher Abkürzung (s. etwa Mhd. Grammatik III, 577f) an; dabei entspricht „bairalem“ dem Sprachraum 2 („ostalemannisch-westbairisch“), „rhfrkhess“ dem Sprachraum 4b („rheinfränkisch-hessisch“) und „thurhess“ dem Sprachraum 5 („hessisch-thüringisch“) von MiGraKo. „mdnd“ steht für die Schreibsprache(n) der hochdeutsch (mitteldeutsch) schreibenden Niederdeutschen.
- Den Schluss bildet die Bezeichnung der Textart: P = Prosa, U = Urkunden, V = Verstext. Im Falle einiger großer Verstexte steht stattdessen die Textsigle: „Rol“ = Pfaffe Konrad, ‚Rolandslied‘, Hs. P; „Kchr“ = ‚Kaiserchronik‘; „Herb“ = Herbort v. Fritzlar, ‚Liet von Troye‘, Hs. H. Die beiden letzten Texte müssen wegen ihres großen Umfangs in ANNIS in zwei gruppenfüllende Teiltexte aufgespalten werden: R12_2-bair-V_Kchr1 und R12_2-bair-V_Kchr1; R14_1-md-ofrk-V_Herb1 und R14_1-md-ofrk-V_Herb2.

Aus der folgenden Übersicht ergibt sich, welche eReM-Gruppen welchen MiGraKo-Gruppen korrespondieren und daher zu deren Erweiterung herangezogen werden können:

MiGraKo-Gruppe	Texte	Tokens	ReM-Gruppe	Texte	Tokens
11_2-12_1-obd-PV-G	7	38198	11-12_1-obd-PV-X	39	22405
			11-12_1-rhfrhess-PV-X	5	832
			12_2-obd-PV-X	9	6215
12_2-bair-PV-G	4	37046	12_2-bair-P-X	22	40104
			12_2-bair-V_1-X	11	45668
			12_2-bair-V_Kchr1-X	4	46560
			12_2-bair-V_Kchr2-X	1	49757
			12_2-bair-V_2-X	10	42798
			12_2-bair-V_3-X	8	33868
			12_2-bair-V_Rol-X	1	44050
12_2-bairalem-PV-G	3	27002	12_2-bairalem-PV-X	7	19621
12_2-alem-PV-G	4	25217	12_2-alem-PV-X	10	8290
12_2-wmd-PV-G	4	17951	12_2-wmd-PVU-X	18	24413
12_2-thurhess-PV-G	2	10653	12_2-omd_ofrk-PV-X	7	2169
13_1-obd-V-G	4	54281	13_1-obd-PV-X	9	5382
13_1-bair-PV-G	4	45056	13_1-bair-P-X	15	32401
			13_1-bair-V-X	17	32529
13_1-bairalem-PV-G	2	27811	13_1-bairalem-PV-X	7	10413
13_1-alem-PV-G	4	25927	13_1-alem-PV-X	7	19236
13_1-wmd-PV-G	3	34728	13_1-wmd-PV-X	15	15598
13_1-thurhess-PV-G	3	48771	13_1-thurhess-PV-X	7	33335
			12-13_1-mdnd-PV-X	4	30264
			12-13_1-mdnd-V-X	5	31619
13_2-bair-PV-G	4	50349	13_2-bair-P-X	4	26389
			13_2-bair-V-X	2	23231
13_2-bairalem-PV-G	4	47326	13_2-bairalem-PV-X	2	8685
13_2-alem-PU-G	3	45550	13_2-alem-PV-X	3	11851
13_2-alem-V-G	2	24787			
13_2-mfrk-PUV-G	5	42455	13_2-mfrk-PU-X	2	47400
13_2-rhfrhess-PV-G	3	34295	13_2-md-V-X	9	29889
13_2-omd-PV-G	3	33079			
14_1-bair-PUV-G	4	43775	14_1-obd-PV-X	7	37336
14_1-bairalem-PUV-G	4	56729			
14_1-alem-PU-G	2	34038			
14_1-alem-V-G	2	29157			
14_1-mfrk-PUV-G	5	55047	14_1-mfrk-P-X	6	26427
			14_1-mfrk-U-X	1	48365
14_1-rhfrhess-PU-G	3	42219	14_1-rhfrhess-PV-X	4	29044
14_1-rhfrhess-V-G	3	47758			
14_1-omd-PU-G	3	34530	14_1-omd-PV-X	5	23781
14_1-omd-V-G	3	44335			
			14_1-md-ofrk-V_Herb1-X	1	44506
			14_1-md-ofrk-V_Herb2-X	2	44329
14_1-ofrk-PU-G	4	41259	14_1-ofrk-P-X	3	21659
14_1-ofrk-V-G	2	18376			

3. Transkription

Grundsätzlich war das Ziel, alle Texte so handschriftengetreu zu erfassen, dass alle linguistisch relevanten Merkmale eindeutig abgebildet sind. Grundlage waren dafür in aller Regel Abbildungen der Handschrift. Ausnahmen davon bilden einerseits Texte, deren Handschrift verschollen oder zerstört ist und die daher nur in mehr oder weniger handschriftengetreuen Abdrucken des 19. Jh. vorliegen; andererseits Texte, für die Handschriftenabbildungen aus unterschiedlichen Gründen (bislang) nicht zu beschaffen waren und für die daher gleichfalls auf Handschriftenabdrucke zurückgegriffen wurde. Diese Ausnahmen sind in ANNIS in den Meta-Annotations unter „notes-transcription“ jeweils vermerkt.

Vor allem für die Transkription der MiGraKo-Texte in den 1990er Jahren musste meist mit Schwarzweißfilmen bzw. mit Rückvergößerungen von ihnen in unterschiedlicher Güte gearbeitet. In einigen Fällen war die Abbildungsqualität so schlecht, dass vieles nur unzureichend oder gar nicht lesbar war. Inzwischen sind vielfach qualitativ hochwertige Farbabbildungen online verfügbar, sodass eine erneute Kollationierung dringend erwünscht wäre. Auch auf diese Fälle wird in den Metadaten unter „notes-transcription“ hingewiesen.

4. Annotationsumfang und Annotationsverfahren

Im folgenden Dokument wird unterschieden zwischen den Annotationen, die im Korpus selbst („ReM“) vorgenommen wurden, und den Annotationen, wie sie in ANNIS umgesetzt und dort verfügbar und abfragbar sind. Die Annotationen in ANNIS stellen eine Teilmenge der Annotationen in ReM dar. Möchte man die Gesamtheit der Annotationen nutzen, kann man das ReM-Korpus in einem XML-Austauschformat herunterladen (aktuell: CorA-XML, in Arbeit: TEI-konformes XML).

4.1 Annotationsumfang

Die Annotation ist rein tokenbezogen. Jedem Token (Wortform) werden zugeordnet (zu weiteren Erläuterungen s. die nachfolgenden Abschnitte):

- das Lemma (mit speziellen Regelungen für Pronominaladverbien und Partikelverben, in ANNIS in den Ebenen `lemma` und `lemmaLemma`),
- die Wortart/POS (ANNIS: `pos` und `posLemma`),
- beim Verb die Flexionsklasse, beim Substantiv ggf. Hinweise zum Pluralumlaut (ANNIS: `inflectionClass` und `inflectionClassLemma`),
- bei den Flektierbaren die Angaben zur Flexion (ANNIS: `inflection`);
- ferner die „normalmhd.“ Version („Normalform“) der handschriftlichen Wortform (ANNIS: `norm`)
- und im gesamten MiGraKo und einem Teil von eReM die graphophonemische Wortformanalyse (ANNIS: `char_align`);
- außerdem ist in ReM die genaue Stellenreferenz jedes Tokens annotiert. In ANNIS wird die Stellenreferenz pro Zeile angegeben (`reference`).

Als wesentliche Neuerung wurde im Rahmen des Projekts „Referenzkorpus Mittelhochdeutsch“ und der Entwicklung von HiTS die Trennung von *allgemeinem* Lemma + Wortart

(`lemmaLemma`, `posLemma`) und *belegspezifischem* Lemma + Wortart (`lemma`, `pos`) eingeführt (vgl. Dipper, et al. (2013) und Abschnitt 7). Bei den MiGraKo-Texten und den aus den Köln-Bonner Mitteldeutsch-Korpora stammenden Texten von eReM wurde diese Trennung nachgerüstet.

In Kurzkommentaren (die bislang noch nicht nach ANNIS importiert wurden) wird ggf. auf kodikologische Besonderheiten, (vermutliche) Überlieferungsfehler oder Annotationsprobleme hingewiesen.

4.2 Annotationsverfahren

Alle Teilkopora von ReM sind mit einem Programmpaket halbautomatisch annotiert worden, das Thomas Klein ab 1986 in Köln entwickelt und später mehrfach erweitert und verbessert hat (Klein, 1991; 2001). Das Verfahren besteht im Wesentlichen darin, dass den Wortformen eines präeditierten (d.h. tokenisierten und interpungierten) Textes automatisch Vorschläge zur Lemmatisierung und grammatischen Annotation zugeordnet werden, die nach einer statistischen Gewichtung geordnet sind. Anschließend muss manuell geprüft werden, ob das Präferenzangebot dieser Vorschläge zutrifft oder durch einen anderen Vorschlag zu ersetzen ist.

Danach werden interaktive Scripte zur Einfügung weiterer Markierungen angewendet. Sodann wird aus Lemma, grammatischer Annotation und handschriftlicher Wortform eine „Normalform“ erzeugt (in ANNIS, `norm`), z.B. (‹é› steht für einen unabgeschwächten Nebensilbenvokal, hier für ‹o› und ‹a›):

(1)

`tok_dipl`	uuare
`norm`	wære
`lemma`	wësen
`inflection`	Subj.Past.Sg.3

(2)

`tok_dipl`	chosota
`norm`	kôsété
`lemma`	kôsen
`inflection`	*.Past.Sg.3

Aus dieser Normalform und der handschriftlichen Wortform wird schließlich interaktiv die graphophonemische Analyse erstellt, z.B. für die beiden obigen Beispiele |uu=w|a=æ|r=r|e=e| und |ch=k|o=ô|s=s|o=é|t=t|a=é|. Die graphophonemische Analyse liegt zur Zeit nur für alle MiGraKo-Texte und für eine kleine Zahl von eReM-Texten vor. Sie wird in ANNIS als `char_align`-Ebene angezeigt.

Erst sekundär wurden die ReM-Texte zum Zweck des ANNIS-Imports durch ein heuristisches Script von Thomas Klein in XML-Dateien mit HiTS-Tags überführt, und zwar – soweit bis jetzt zu sehen – korrekt. Nur einige Tags, für die es in HiTS keine Entsprechung gibt, blieben von

der Überführung ausgenommen. Ein Problem stellt u.a. die unterspezifizierte interne Annotation (s. unten Abschnitt 8.2) dort dar, wo in HiTS eine spezifizierte gefordert oder zumindest erwünscht ist.

5. Tokenisierung

In ReM wird zwischen der handschriftlichen Tokenisierung (`tok_dipl`) und einer modernen Erwartungen entsprechenden Tokenisierung (`tok_anno`) unterschieden. Insofern als diese Tokenisierungen auseinander gehen, wird handschriftliche Getrenntschreibung von Wortteilen oder Zusammenschreibung von Wörtern auf der `tok_dipl`-Ebene in `tok_anno` aufgehoben. Die Tokens auf der Ebene `tok_anno` dienen als Referenzpunkte für alle weiteren linguistischen Annotationen (z.B. Lemma, Morphologie, Wortart), während die Stellenreferenz (Angaben zur Seite, Zeile, Spalte etc.) an den diplomatischen Tokens verankert ist.

Die Art der Differenz zwischen der diplomatischen und der modernisierten Tokenisierung wird auf der `tokenization`-Ebene festgehalten. Die folgenden Fälle werden unterschieden:

(3) Multiverbierung mit Spatium

Wenn ein handschriftliches Token modern mehreren durch Spatien getrennte Tokens entspricht. Jede Wortform im `tok_anno` wird mit MS1, MS2, … versehen.

`tok_dipl`	indem		lande
`tok_anno`	in	dem	lande
`tokenization`	MS1	MS2	

(4) Univerbierung mit Spatium

Mehrere handschriftliche Wortformen (durch Spatien getrennt), die modern zusammen geschrieben werden.

`tok_dipl`	der	burger	meister
`tok_anno`	der	burgermeister	
`tokenization`		US	

(5) Univerbierung am Zeilenende

Wenn beim Zeilenumbruch die Zusammengehörigkeit zweier Wortformen nicht handschriftlich mit einem Strich gekennzeichnet wird oder dieses Trennzeichen nicht erkennbar ist.

`tok_dipl`	under ↵ tan
`tok_anno`	undertan
`tokenization`	UL

(6) Multiverbierung am Zeilenende

Wenn in der Handschrift ein Bindestrich am Zeilenende steht, um die Zusammengehörigkeit zweier Wortformen zu kennzeichnen, die nach moderner Auffassung zwei separate Wortformen darstellten. Wortformen werden mit ML1, ML2, … versehen.

tok_dipl	de= ↵	bogen
tok_anno	de	bogen
tokenization	ML1	ML2

Klitika werden auch bei handschriftlicher Zusammenschreibung als eigene Tokens gewertet:

(7)

tok_dipl	er	ensagetez		
tok_anno	er	en	saget	ez
tokenization		MS1	MS2	MS3

Gemischte Fälle von Uni- und Multiverbierung sind möglich, hier werden die entsprechenden Tokens mit mehreren Tags versehen:

(8)

tok_dipl	be	durfeter
tok_anno	bedurfet	er
tokenization	US MS1	MS2

6. Interpunktion

Die Annotationsebene punc enthält Informationen zum Satztyp, die in modernen Texten durch Interpunktion kodiert würde. Handschriftliche Interpunktion wird stets abgetrennt, und wenn dieses Zeichen eine satzbeendende Funktion hat, wird es mit dem punc-Attribut „$E“ versehen. Tags, die den Satztyp angeben, sind am letzten Wort im Satz oder Teilsatz angefügt, ob handschriftliche Interpunktion vorhanden ist oder nicht, z.B. *Von gold ein uv̊rſpan harte wol Geſmidit un̄ edilir ſteine uol*/**DE** *Daz im argiſ niht geſchach*/**DE** *./***$E** M301, 6a,167. In ANNIS wird dieses Beispiel so dargestellt:

	7a,1					
uol	Daz	im	argiſ	niht	geſchach	.
uol	Daz	im	argis	niht	geschach	.
voll	dazz	im	arges	niht	ge-schah	
--	--	--	--	--	--	
ADJN	KOUS	PPER	NA	NA	VVFIN	
ADJ	KO	PPER	NA	NA	VV	
voll	dazz	ër	arg	niht	ge-schëhen	
212331000	29484000	40380000	8799000	121449000	55767000	
voll	dazz	ër	arg	niht	ge-schëhen	
Pos.*.Gen.Pl.0	--	Masc.Dat.Sg.3	Gen.Sg	Nom.Sg	Ind.Past.Sg.3	
--	--	--	st.Neut	st.Neut	st5	
--	--	--	st.Neut	st.Neut	st5	
|u=v|o=o|l=ll|	|d=d|a=a|z=zz|	|i=i|m=m|	|a=a|r=r|g=g|i=e|s=s|	|n=n|i=i|h=h|t=t|	|g=g|e=e|=-|sch=sch|a=a|ch=h|	
DE					DE	$E

Neben dem einfachen Punkt (.), werden in ReM auch die handschriftlichen (`tok_dipl`) Formen · (halbhoher Punkt) und ⁝ (punctus elevatus) unterschieden. Alle drei Formen werden auf `tok_anno`-Ebene als normale Punkte (.) dargestellt.

Tags für Satzgrenzen: („E“ steht jeweils für „Ende des Satzes“)

DE	Ende eines Deklarativsatzes
IE	Ende eines Imperativsatzes
EE	Ende eines Exklamativsatzes
QE	Ende eines Interrogativsatzes

Sonstige Tags für Interpunktion:

S*	Markiert die Grenzen von Teilsätzen eines Satzgefüges sowie von Links- und Rechtsversetzungen. Z.B. *Min uil lieben*/**S*** *unſ ſaget daz heilige eu(angel)*ium/**S*** . *daz wir hivte gileſen haben*/**S*** . *wie unſer herre zeinen ſtunden kom zů dem mere an galilee lande* M165, 27v,17; *wariu uogitinne*/**S*** *nu bihůtte . dine ſcalche hie unde ouch da*/**S*** . *sancta maria* . M107G, 332.
N*	Markiert Elemente innerhalb einer Aufzählung oder Appositionsreihe. (Beispiel s. NE)
NE	Markiert das Ende einer Aufzählung bzw. Appositionsreihe. Beginnt nach der Apposition ein anderer Teilsatz oder eine Rechtsversetzung, so wird die Grenze mit S* anstatt NE markiert. Z.B. *Sancte ſilueſter*/**N*** . *der hailige man*/**NE** . *hiez hieremiam da uŏr tragen* M121V, 8884.

7. Lemmatisierung

7.1. Allgemeines

Es wird unterschieden zwischen dem allgemeinen Lemma (`lemmaLemma`) und dem belegspezifischem Lemma (`lemma`). In den meisten Fällen sind beide identisch. Bei Pronominaladverbien und Partikelverben erhält das Beleg-Lemma die entsprechenden Zusätze (z.B. kann dem allgemeinen Lemma *dâr* das Beleg-Lemma *dâr/+hin(e)* entsprechen, s. 7.4 bzw. 7.5).

Die Gestaltung der Lemmaansätze und die Bestimmung der grammatischen Kategorien richtet sich meist nach dem jeweils entsprechenden Ansatz in LEXERs Handwörterbuch. Zusätzlich gelten die im Folgenden beschriebenen Regelungen, die die Unterschiede zwischen den Lemmaansätzen bei LEXER und in MiGraKo/ReM erklären.

- Nach dem Vorbild des LEXER-Ansatzes werden Präfixe, Suffixe und Kompositionsglieder durch Segmentierungsstriche abgetrennt. Diese morphologische Segmentierung wird konsequent auch bei vielgliedrigen Zusammensetzungen durchgeführt, während LEXER etwa bei mehrfacher Präfigierung nur das erste Präfix segmentiert, z.B. LEXER *un-gesellеschaft* – ReM *un-ge-sèlle-schaft*.
- Die Bestandteile von Komposita werden im Lemmaansatz möglichst so geschrieben wie die jeweiligen Simplicia; so wird z. B. nach dem Simplex *wünne* auch *wünne-brunne* angesetzt (gegen *wunne-brunne* bei LEXER).
- Feste Fügungen mit *in-* oder *en-* (z. B. *en-triuwen*, *en-gègen*) werden als ein Lemma angesehen, wenn es in LEXERs „Handwörterbuch“ einen entsprechenden Ansatz oder wenigstens einen entsprechenden Verweis gibt.
- Verbindungen aus *aller* und dem Superlativ eines Adjektivs oder Adverbs werden als ein Wort angesetzt (z. B. *aller-schoènest*).
- Adverbien mit verschiedenen erstarrten Flexionsendungen oder univerbierte adverbiale Syntagmen sind jeweils als ein eigenes Lemma angesetzt (z. B. *tages*, *all-ze-hande*).

7.2. Lautliche Besonderheiten

- Grundlage für den Lemmaansatz ist der frühmhd. Sprachstand, nicht der des „klassischen“ Mhd. Lachmann’scher Prägung. Dies gilt insbesondere für die Erhaltung von Schwa (teils auch *-i-*) und das noch nicht lenisierte ahd.-frühmhd. *-nt-*; daher z.B. *tür(e)* ‚Tür‘, *hèr(e)-bërge*, *mèn(ni)sche* (zur Klammerung s. u. 7.3), *binten* ‚binden‘, *unte* ‚und‘. Bei Unsicherheit, ob bei Suchanfragen *nt* oder *nd* zu wählen ist, sollte man bei der ANNIS-Suche von regulären Ausdrücken Gebrauch machen, z.B. `lemma=/bin[dt]en/`.
- Ebenso wird die Unterscheidung von ahd. *gg* (< westgerm. **gg*) und *ck* (< westgerm. **kk*) beibehalten, z.B. *mügge* ‚Mücke‘ – *stücke* ‚Stück‘, *ègge* ‚Ecke, Schwertschneide‘ – *dècke* ‚Decke‘.
- *e*-Laute:

  Bei den mhd. kurzen *e*-Lauten unterscheidet LEXER (und MWB) nur ‹ë› für /ɛ/ (< westgerm. **e*) und ‹e› für Umlaut-e und Schwa. In ReM wird das Umlaut-*e* dagegen durch ‹è› bezeichnet und dadurch von ‹e› für Schwa unterschieden, z.B. LEXER *gegen-rede* –

ReM *gègen-rède*, LEXER *her-bërge* – ReM *hèr(e)-bërge*. Zwischen Primär- und Sekundärumlaut-*e* wird in ReM wie bei LEXER (und MWB) nicht unterschieden, z.B. *mèhtig* (nicht *mähtig*), *gèrwen* (nicht *gärwen*).

- ‹è› wird außerdem in den Diphthongen mhd. *iè*, *üè* und in ‹oè› für mhd. *œ* verwendet: *bièten*, *güète*, *schoène*.
- Als Deckelzeichen für ein *e* unklarer Qualität wird *É* verwendet, also ein großes, akutiertes ‹e›. Dieses *É* wird vor allem für in Lehnwörtern gebraucht, z.B. *tÉmpel* ‚Tempel'.
- Silbenauslaut

Die ReM-Lemmata sind anders als bei LEXER ohne Auslautverhärtung und Geminatenkürzung angesetzt:

Lexer	ReM
walt, -des	wald
tac, -ges	tag
wîp, -bes	wîb
brief, -ves	brièv
brief-buoch	brièv-buoch
schuoch, -hes	schuoh
durch	durh
sin, -nnes	sinn
blic, -ckes	blick

- Auch der Auslaut von *-wa-/-wō*-Stämmen wird mit *-w* angesetzt, *varw* adj, *sêw* m. (nicht *var*, *vare* bzw. *sê*)
- Das Adjektivsuffix *-ig* (bei LEXER meist *–ec*) erscheint grundsätzlich – im Inlaut wie im Auslaut – als *-ig*, z. B. *man**ig*** statt *manec*, *man**ig**-valt* statt *manec-valt*, *man**ig**-valt**ig**-hèit* statt *manec-valtecheit*.
- Vor den Suffixen *-inne*, *-în* (Adj.), *-ede*, *-lîn* und *-chîn* wird das Lemma grundsätzlich mit Umlaut angesetzt, vor *-ære*, *-lich* und *-lîche* grundsätzlich ohne Umlaut. Vor den Suffixen *-ig*, *-nisse* richtet sich der Umlaut des Lemmaansatzes nach dem entsprechenden Ansatz bei Lexer.

7.3.Zusammenfassung von Lemmavarianten

- Vielfach setzt LEXER bei einem Lemma Varianten mit und ohne auslautendes /-e/ an, so stets im Falle der mhd. Apokope nach Kurzvokal + Liquid (z. B. „*tür*, *türe*"). Besonders bei mitteldeutschen und frühen oberdeutschen Texten muss dem – hier sehr oft (noch) erhaltenen – finalen /-e/ Rechnung getragen werden; dies geschieht, indem das /-e/ im Lemmaansatz in runden Klammern angefügt wird (etwa *tür(e)*).
- Auch sonst können Lemmavarianten durch Rundklammerung zusammengefasst sein, z.B. *mèn(ni)sche* für LEXER *mensche*, *mensch*, *mennische*; *iè-mann(d)* für *iè-mann* und *iè-mannd*; *wîs(e)* Adj für *wîs* (*a*-Stamm) und *wîse* (*ja*-Stamm). Es ist möglich, z.B. vom allgemeinen

Lemma *wîs(e)* das belegspezifische *wîs(e)* bzw. *wîs* zu unterscheiden. Von dieser Möglichkeit ist bislang aber nur beschränkt Gebrauch gemacht worden, und zwar auch deshalb, weil die Unterscheidung von allgemeinem und belegspezifischem Lemma in MiGraKo und nicht wenigen eReM-Texten erst nachträglich eingeführt wurde.

- Gibt es bei einem Lemma rein lautliche (also nicht die Wortbildung betreffende) *Varianten*, so gilt in der Regel der erste LEXER-Lemmaansatz, z. B. bei Varianten mit und ohne Umlaut: LEXER „*mügen, mugen*“ → ReM *mügen*.
- Feminina auf *-inne*/*-în* werden immer auf *-inne* angesetzt; nur die Normalform wird in einem späteren Arbeitsschritt angepasst, wenn die handschriftliche Form auf *-în* ausgeht.
- Komposita mit *voll-*, *volle-* oder *vollen-* als erstem Kompositionsglied werden stets mit *voll-* angesetzt; Komposita mit *all-*, *alle-*, *allen-* oder *allent-* stets mit *all-*. Die Normalformen sind dagegen der jeweiligen handschriftlichen Wortform angepasst.
- Adverbien auf *-lîche* oder *-lîchen* sind unter dem Ansatz auf *-lîche* zusammengefasst.
- Für die dialektalen Varianten von ‘oder’ (*ove*/*of*, *obe*, *ave(r)*, *ald(er)* usw.) ist stets *oder* als Lemma angesetzt. Erst in der Normalform erscheint die dialektale Variante.
- Einheitlicher Ansatz des Präfixes *ent-* unabhängig vom folgenden Konsonanten, z.B. *ent-bër(e)n*, *ent-gëlten*, *ent-vinden* (LEXER *en-bërn*, *en-gëlten*, *enphinden*).
- Nomina agentis mit dem Suffix ahd. *-āri*, die Lexer teils mit *-ære*, teils mit *-er* anführt, werden einheitlich mit *-ære* angesetzt. Einwohnernamen haben dagegen das Suffix *-er(e)*, z.B. *burger(e)* ‚Bürger‘, *Rômer(e)* ‚Römer‘

7.4.Pronominaladverbien

Wegen möglicher Distanzstellung sind der adverbiale und der präpositionale Teil von Pronominaladverbien als separate Tokens angesetzt. Ihre Zusammengehörigkeit wird durch entsprechende Zusätze im belegspezifischen Lemma (`lemma`) verdeutlicht, wobei zwischen Kontakt- und Distanzstellung (markiert durch „.“) unterschieden wird:

		`lemmaLemma`	`lemma`
a) Kontaktstellung	Adverb	dâr	dâr/+hin(e)
	Präposition	hin(e)	hin(e)/dâr+
b) Distanzstellung	Adverb	dâr	dâr/.+hin(e)
	Präposition	hin(e)	hin(e)/dâr.+

(a) ſwa er daz ſin lieb weiz, **da hin** chert er ... ſin gemv̊te M214, 62v,6

(b) *daz vêlt*, **da** *iz* **hin** *lǒfet* M214, 10v,20

7.5.Partikelverben

Wegen möglicher Distanzstellung sind Partikel und Basisverb von Partikelverben als separate Tokens behandelt worden. Die Zusammengehörigkeit beider Elemente wird durch Zusätze beim belegspezifischen Lemma (`lemma`) verdeutlicht, z.B. bei *ane sëhen*:

	`lemmaLemma`	`lemma`

Partikel	ane	ane/+sëhen
Basisverb	sëhen	sëhen/ane+

Diese Bezeichnungsweise gilt für MiGraKo und in eReM für die aus den Köln-Bonner Mitteldeutsch-Korpora aufgenommenen Texte. In den übrigen eReM-Texten ist darüber hinaus markiert, ob die Partikel dem Basisverb vorangeht (>) oder umgekehrt (<) und ob Kontakt- oder Distanzstellung vorliegt. Distanzstellung wird durch einen Punkt bezeichnet:

			`lemmaLemma`	`lemma`
(1) Partikel geht voran	a) Kontaktstellung	Partikel	ane	ane/>+sëhen
		Basisverb	sëhen	sëhen/ane>+
	b) Distanzstellung	Partikel	ane	ane/>.+sëhen
		Basisverb	sëhen	sëhen/ane>.+
(2) Verb geht voran	a) Kontaktstellung	Basisverb	sëhen	sëhen/ane<+
		Partikel	ane	ane/<+sëhen
	b) Distanzstellung	Basisverb	sëhen	sëhen/ane<.+
		Partikel	ane	ane/<.+sëhen

(1a) *uil gůtlichen er in* **ane ſach** M024, 1091

(1b) *der rote apfphel, der iſt liepliche* **ane** *ze* **ſehenne** M113y, 33r,17

(2a) **ſihe an** *die diemůte min!* M193, 22v,14

(2b) **ſich** *unſer diemůt* **an**! M119, 734

In ReM sind die Partikel und das zugehörige Basisverb zusätzlich durch den Pointer markiert, bei denen der Zeiger auf die ID des jeweiligen anderen Elements verweist (in ANNIS aktuell noch nicht importiert).

7.6.Lateinisches

Lateinische (oder sonst fremdsprachliche) Wörter sind durch „FM“ als allgemeine Wortart (`posLemma`) gekennzeichnet. Sie haben als allgemeines Lemma (`lemmaLemma`) meist „[!]“. Wenn sie zu einem teilweise deutschen Satz gehören, werden sie im Übrigen wie deutsche Wörter annotiert. Stehen sie dagegen in rein lateinischem Kontext, so sollten sie als belegspezifisches Lemma (`lemma`, und als Normalform, `norm`) die normalisierte handschriftliche Wortform erhalten, ansonsten aber nicht annotiert sein. Diese Regelung ist aber in eReM nur sehr unsystematisch durchgeführt worden. Häufig stehen hier im Zuge der Annotationsprozedur automatisch erzeugte Vergleichsformen, die zumal als lateinische Lemmata ganz unangemessen sind.

8. Grammatische Annotation (Wortart, Flexion)

8.1. Vorbemerkung

Alle ReM-Texte sind primär in einem Tag-System annotiert worden, das 1986 von Thomas Klein und Joachim Bumke für das Kölner Mitteldeutsch-Projekt entworfen und in Bonn im Projekt Mhd. Grammatik modifiziert und erweitert wurde: So kam im Bereich der morphosyntaktischen Annotation insbesondere die Unterscheidung von Pronomen und Artikel bei *dër* und *ein* und die Markierung von substantivischen und attributiv nachgestellten Adjektiven und Pronomen hinzu.

8.2. Wortart (POS)

Bei Substantiven besteht die Wortartbestimmung (POS) in der internen Annotation nur aus der Angabe des Genus (und ggf. einem Hinweis zum Pluralumlaut, s. unten). Auf eine Flexionsklassenbestimmung ist dagegen grundsätzlich verzichtet worden, weil sie sich zum einen aus der Gestalt des Lemmas ergibt: Auf Kons. endende Substantive gehören fast ausnahmslos zur „starken" Deklination. Zum andern schwanken insbesondere die Feminina auf *-e* in dialektal unterschiedlichem Ausmaß zwischen starker und schwacher Deklination. Eine Zuweisung von „st" (stark) und „wk" (schwach) ist bei ihnen daher nur belegspezifisch möglich. – Zur nachträglichen Einfügung von Flexionsklassenbestimmungen in HiTS s. unten 8.3.

Bei Substantiven mit regelmäßigem Pluralumlaut wird der Genusangabe <u> hinzugefügt, bei nur fakultativem Pluralumlaut <(u)>, z. B. *gast* m<u> (Plural *gèste)*; *hant* f<(u)> (Plural *hande* ~ *hènde*).

In ANNIS wird generell zwischen allgemeiner (`posLemma`) und belegspezifischer (`pos`) Wortart unterschieden, z.B. für eine Form wie *ahtet* könnten die Wortart-Annotationen so belegt werden: `posLemma="VV"` (Verb), `pos="VVFIN"` (finites Verb) (s. Dipper, et al. (2013) über HiTS-Tags). Die folgende Tabelle zeigt die in REM genutzten POS-Tags der Beleg-Ebene. Die POS-Tags der Lemma-Ebene sind im Allgemeinen gekürzte Varianten der Beleg-POS-Tags.

	POS	**Beschreibung**
Adjektiv	ADJA	Adjektiv, attributiv, vorangestellt
	ADJD	Adjektiv, prädikativ
	ADJN	Adjektiv, attributiv, nachgestellt
	ADJS	Adjektiv, substituierend
Adposition	APPR	Präposition
Adverb	AVD	Adverb
	AVD-KO*	Adverb oder Konjunktion
	AVG	Relativadverb, generalisierend
	AVW	Adverb, interrogativ
Kardinalzahl	CARDA	Kardinalzahl, attributiv, vorangestellt
	CARDD	Kardinalzahl, prädikativ

	CARDN	Kardinalzahl, attributiv, nachgestellt
	CARDS	Kardinalzahl, substituierend
Determiner	DDA	Determinativ, definit, attributiv, vorangestellt
	DDART	Determinativ, definit, artikelartig
	DDD	Determinativ, definit/demonstrativ, prädikativ
	DDN	Determinativ, definit/demonstrativ, attributiv, nachgestellt
	DDS	Determinativ, definit/demonstrativ, substituierend
	DGA	Determinativ, generalisierend, attributiv, vorangestellt
	DGN	Determinativ, generalisierend, attributiv, nachgestellt
	DGS	Determinativ, generalisierend, substituierend
	DIA	Determinativ, indefinit, attributiv, vorangestellt
	DIART	Determinativ, indefinit, artikelartig
	DID	Determinativ, indefinit, prädikativ
	DIN	Determinativ, indefinit, attributiv, nachgestellt
	DIS	Determinativ, indefinit, substituierend
	DPOSA	Determinativ, possessiv, attributiv, vorangestellt
	DPOSD	Determinativ, possessiv, prädikativ
	DPOSN	Determinativ, possessiv, attributiv, nachgestellt
	DPOSS	Determinativ, possessiv, substituierend
	DRELS	Determinativ, relativisch, substituierend
	DWA	Determinativ, interrogativ, attributiv, vorangestellt
	DWD	Determinativ, interrogativ, prädikativ
	DWS	Determinativ, interrogativ, substituierend
Fremdsprachl.	FM	Fremdsprachliches Material
Interjektion	ITJ	Interjektion
Konjunktion	KO*	Konjunktion, neben- oder unterordnend
	KOKOM	Vergleichspartikel
	KON	Konjunktion, nebenordnend
	KOUS	Konjunktion, unterordnend
Nomen	NA	Nomen appelativum
	NE	Eigenname
Pronominaladverb	PAVAP	Pronominaladverb, präpositionaler Teil
	PAVD	Pronominaladverb, pronominaler Teil

	PAVG	Pronominaladverb, pronominaler Teil, generalisierend
	PAVW	Pronominaladverb, pronominaler Teil, interrogativ
Pronomen	PG	Pronomen, generalisierend
	PI	Pronomen, indefinit
	PPER	Pronomen, personal, irreflexiv
	PRF	Pronomen, personal, reflexiv
	PW	Pronomen, interrogativ
Partikel	PTKA	Partikel bei Adjektiv oder Adverb
	PTKANT	Antwortpartikel
	PTKNEG	Negationspartikel
	PTKVZ	Verbzusatz
Verb	VAFIN	Auxiliar, finit
	VAIMP	Auxiliar, imperativ
	VAINF	Auxiliar, infinitiv
	VAPP	Auxiliar, Partizip Präteritum, im Verbalkomplex
	VAPS	Auxiliar, Partizip Präsens, im Verbalkomplex
	VMFIN	Modalverb, finit
	VMIMP	Modalverb, imperativ
	VMINF	Modalverb, infinitiv
	VMPP	Modalverb, Partizip Präteritum, im Verbalkomplex
	VMPS	Modalverb, Partizip Präsens, im Verbalkomplex
	VVFIN	Vollverb, finit
	VVIMP	Vollverb, imperativ
	VVINF	Vollverb, infinitiv
	VVPP	Vollverb, Partizip Präteritum, im Verbalkomplex
	VVPS	Vollverb, Partizip Präsens, im Verbalkomplex
Interpunktion	$_	originale Interpunktion

Während an der unterspezifizierten Annotation des Modus auch in HiTS nichts zu ändern ist, werden die Flexionsklassenbestimmungen „st“ (stark), „wk“ (schwach) und „*“ (stark oder schwach) bei Substantiven nach folgender Regel zugewiesen:

A. Lemmabezogen (`inflectionClassLemma`)

1. Alle auf Konsonant oder Langvokal endenden Lemmata sind „st“.
2. Bei den auf *-e* oder *-(e)* endenden Lemmata wird nach Genus unterschieden:
   a. Das Genus ist Masc. und das Lemma
      i. endet auf *-ære* oder *-er(e)*, dann „st“,

ii. gehört zu der Gruppe (oder enthält eines ihrer Lemmata als letztes Kompositionsglied) *kæse*, *wèize*, *mütte*, *wite*, *wine* oder *mëte*, dann „st“,
iii. gehört zu der Gruppe der ca. 120 in MiGraKo stets schwach flektierenden Masc. auf *-e* oder *-(e)* (oder enthält eines dieser Lemmata als letztes Kompositionsglied), dann „wk“,
iv. alle sonstigen masc. Lemmata auf *-e* oder *-(e)* sind von unbestimmter Flexionsklasse: „*“.

b. Das Genus ist Neut. und das Lemma
i. gehört zu der Gruppe (oder enthält eines ihrer Lemmata als letztes Kompositionsglied) *hërze*, *ouge*, *ôre*, *wange*, *junge* (‚Tierjunges‘) dann „wk“;
ii. sonst ist es „st“.

c. Alle fem. Lemmata auf *-e* oder *-(e)* sind von unbestimmter Flexionsklasse: „*“.

B. Belegbezogen (`inflectionClass`).

1. Bei den Substantiven von unbestimmter Flexionsklasse, also den Masc. nach A 2.a.iv und den Fem. auf *-e* oder *-(e)* (A 2.c) bleiben auch die Einzelbelege im Nom.Sg. und Dat.Pl., bei Fem. auch im Gen.Pl., unbestimmt.
2. Für die übrigen Kasus gilt:
   a. Endet die Wortform mit dem Flexiv *-(e)n*, dann „wk“,
   b. sonst „st“.

   z.B. *rëde* *.Fem: *rëde* Nom.Sg, *.Fem; *rëden* Dat.Sg, wk.Fem; *rëde* Dat.Sg, st.Fem; *rëden* Gen.Pl, *.Fem.

Die meisten Substantive erhalten also in der ANNIS-Version von ReM eine generelle Flexionsklassenbestimmung (wobei es in Einzelfällen aufgrund der obigen Regel auch zu Fehlzuweisungen kommen kann). Ausgenommen bleiben die Feminina auf *-e* oder *-(e)* und eine Anzahl von Masculina auf *-e* oder *-(e)*, sowohl alte *ja*- als auch *-(j)an*-Stämme, die seltener belegt sind und/oder zwischen starker und schwacher Flexion schwanken. Bei ihnen wird die Flexionsweise nur für den Einzelbeleg als „st“ oder „wk“ angezeigt, wenn eine eindeutige Flexionsendung vorliegt.

In HiTS haben Substantive mit obligatorischen Pluralumlaut „stu“, die mit fakultativem Pluralumlaut „st(u)“ als Flexionsklasse (`inflectionClass`).

8.3.Flexion

Flexionsmerkmale werden in ANNIS auf der `inflection`-Ebene zusammengefasst. Bestimmte Flexionsmerkmale sind nur unterspezifiziert annotiert worden. So wird der Modus nur dort angegeben, wo er ausdrucksseitig an der handschriftlichen Wortform markiert ist (z.B. durch eine moduseindeutige Flexionsendung oder durch Umlaut). Ähnlich ist bei der Angabe des Genus von Adjektiven und Pronomen verfahren worden. Im Gen.Pl. und Dat.Pl. ist das Genus hier an der Wortform allein grundsätzlich nicht erkennbar und wird daher nicht angegeben. Dasselbe gilt für die Flexionsweise (stark „st“ – schwach „wk“) im Dat.Pl. nach der Nebensilbenabschwächung, durch die ahd. -). Im Nom.,Akk.Pl. stehen sich im Oberdeutschen *-e* Masc.,Fem. und *-iu* Neut. gegenüber, während es im Mitteldeutschen keine Genusunterscheidung gibt. Daher ist das Genus (Neut.) lediglich bei Vorliegen des Flexivs *-iu* (oder Varianten) bezeichnet, nur in MiGraKo konsequent auch in den übrigen Fällen.

9. Annotationsfehler und Korrekturstand

Wohl kein annotiertes Korpus größeren Umfangs ist völlig fehlerfrei. Die Annotationsfehler können von leichten Verletzungen der Annotationsstandards bis hin zu gravierenden Fehlern in Lemmatisierung und/oder grammatischer Annotation reichen, die in Missverständnissen des zu annotierenden Textes begründet sind. Bei dem hier praktizierten Annotationsverfahren kommt hinzu, dass bei der (ersten) manuellen Korrektur einzelne Fehler im Präferenzangebot leicht übersehen werden, z.B. falscher Kasus bei ansonsten richtigen Angaben (Lemma, Wortart, etc.). Daher ist hier wenigstens eine nochmalige Abschlusskorrektur wichtig. In MiGraKo ist oft sogar noch zweimal nach der Erstkorrektur korrigiert worden. Bei einer Reihe von eReM-Texten war demgegenüber innerhalb der Projektlaufzeit keine Abschlusskorrektur mehr möglich, sodass die Fehlerrate hier zwangsläufig höher ist. Dies ist zum einen daran erkennbar, dass in der Textbezeichnung (document name) „N0“ oder „G0“ statt „N1“ bzw. „G1“ steht (z.B. „M012-N0“); außerdem ist in den Metadaten das Feld „proofreading_by“ leer. Es besteht die Absicht, auch bei diesen Texten noch eine Abschlusskorrektur vorzunehmen.

Literaturverzeichnis

Dipper, S., Donhauser, K., Klein, T., Linde, S., Müller, S., & Wegera, K.-P. (2013). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics, Special Issue, 28*(1), S. 85–137.

Klein, T. (1991). Zur Frage der Korpusbildung und zur computergestützten grammatischen Auswertung mittelhochdeutscher Quellen. In K.-P. Wegera, *Mittelhochdeutsche Grammatik als Aufgabe* (S. 3–23). Berlin: E. Schmidt.

Klein, T. (2001). Vom lemmatisierten Index zur Grammatik. In S. Moser, P. Stahl, W. Wegstein, & N. R. Wolf, *Maschinelle Verarbeitung altdeutscher Texte V* (S. 83–103). Berlin: de Gruyter.

Klein, T., & Bumke, J. (1997). *Wortindex zu hessisch-thüringischen Epen um 1200.* Tübingen: Niemeyer.

Wegera, K.-P. (Hrsg.). (1991). Mittelhochdeutsche Grammatik als Aufgabe. *Zeitschrift für deutsche Philologie, 110.*

Wegera, K.-P. (2000). Grundlagenprobleme einer mittelhochdeutschen Grammatik. In W. Besch, & al. (Hrsg.), *Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (2 Ausg., Bd. 2, S. 1304–1320). Berlin.

Weimann, B. (2012). *Moselfränkisch: Der Konsonantismus anhand der frühesten Urkunden.* Wien: Böhlau.