Annotationshandbuch
Teil 2: Lemmatisierung

Fabian Barteld, Katharina Dreessen,
Sarah Ihden, Ingrid Schröder
(Institut für Germanistik – Universität Hamburg)

Meike Glawe, Verena Kleymann, Norbert Nagel,
Robert Peters, Elmar Schilling
(Germanistisches Institut – Universität Münster)

12. September 2016

Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200-1650)

gefördert durch die

DFG

Inhaltsverzeichnis

1 Grundsätzliches

Es gibt nur eine einzige Lemmatisierungsebene.

Für jede Wortform wird jeweils nur ein Lemma festgelegt, auch wenn die Lemmaliste mehrere (im LBCM oder LW fettgedruckte) Hauptlemmata aufweisen sollte. Hier darf nur lemmatisiert

werden, was in der Lemmaliste steht (d.h. bei dem Annotationswerkzeug CorA vorgeschlagen wird). Für eventuell erforderliche Neuansetzungen oder das Nachtragen von Lemmata siehe Kap. Neuansetzung und Nachtragen von Lemmata.

Grundlage bildet die vom Standort Münster erstellte Lemmaliste (vgl. Kleymann et al. 2015), die für eine effektive Durchsuchbarkeit wie folgt angepasst wurde:

- Alle Quellen-Markierungen wie "*", "°", "°°", "§" etc. wurden entfernt und werden nicht publiziert.
- Alle mit dem Bindestrich vorgenommenen Wortstamm-Markierungen wurden entfernt und werden nicht publiziert.

1.1 Terminologie

- **Hauptlemma**: fettgedruckte Form in den Wörterbüchern; eigentliches Lemma, auf das, sofern Variantenlemmata vorhanden sind, verwiesen wird.
- **Nebenform**: graphische Variante; erscheint im LBCM kursiv gedruckt hinter dem Hauptlemma; wird nicht in der Lemmaliste berücksichtigt.
- **Variantenlemma**: fettgedruckte Form in den Wörterbüchern, die jedoch auf eine andere Form, das Hauptlemma, verweist; wird nicht veröffentlicht.
- **Nestartikel**: Kursiv gedrucktes Hauptlemma in den Wörterbüchern, das im LBCM mit einer Tilde und im LW mit einem Bindestrich vorweg gekennzeichnet ist, innerhalb eines Wörterbucheintrages steht und i.d.R. ein Kompositum darstellt, bspw. `~schilt` unter dem Hauptlemma `kampmê`^{i}`ster`, das folglich als `kampschilt` zu verstehen ist.

1.2 Ausgeschlossene Token

Token, die mit folgenden PoS-Tags annotiert sind, werden **nicht** lemmatisiert und erhalten in spitzen Klammern die Auszeichnung `none`:

POS-Tag	Lemma
OA	<none>
ED	<none>
XY	<none>
$;	<none>
FM	<none>
PTKVZ	<none>

2 Vorgehen in CorA

2.0.1 Kommentierung des Lemmas mit hilfreichen Informationen

Zur Unterstützung des Lemmatisierungsvorganges in CorA werden in die Lemmaliste auch weitere Informationen aus dem Wörterbuch übernommen wie Wortart, Genus und Hauptbedeutung. Diese Angaben werden nicht veröffentlicht.

Informationsangabe bei Eigennamen Bei Eigennamen, die nicht in den Wörterbüchern und folglich nicht in der Lemmaliste erscheinen, entspricht in der Veröffentlichung das Lemma dem Token, z.B. `bertold` oder `Bertholt`. Auf eine Vereinheitlichung, d.h. die Festlegung je eines Lemmas für verschiedene Varianten eines Namens, muss verzichtet werden.

2.0.2 Neuansetzung und Nachtragen von Lemmata

2.0.2.1 Grundsätzliches

Bei den neu in die Lemmaliste aufzunehmenden Lemmata handelt es sich um die folgenden vier Gruppen:

1) Vergessene Lemmata
Lemmata, die zwar im Wörterbuch als Hauptlemmata erscheinen, aus Versehen in der Lemmaliste aber nicht erfasst wurden,
Bsp.: `egget` (Adjektiv, 'schneidend'),

2) komplett neu angesetzte Lemmata
Lemmata, die im Wörterbuch in keiner Form erscheinen, die aber keinem anderen Hauptlemma zugeordnet, sondern als eigenes Lemma ausgewiesen werden sollen,
Bsp.: `bēkerlîn` (Diminutiv, 'Becherlein'),

3) Nestartikel
kursiv gedruckte, durch die Voranstellung einer Tilde im LBCM bzw. eines Bindestrichs im LW gekennzeichnete Lemma, die innerhalb eines Artikels erscheinen, vgl. Kap. Terminologie,
Bsp.: `klôsterkerke` (Kompositum, 'Klosterkirche'),

4) Ableitungen
kursiv gedruckte Lemmata innerhalb des Artikels eines Hauptlemmas, die Ableitungen dieses Hauptlemmas darstellen, z.B. Adjektivadverbien, Substantivierungen, Diminutivformen etc.,
Bsp.: `bēkerêrsche` (abgeleitetes Femininum, 'Frau des Bechermachers').

2.0.2.2 Zu Beachtendes

Soll ein Lemma neu eingetragen werden, zu dem ein (mit Ausname der Längenbezeichnungen) homographes Lemma existiert, wird dieses analog zu den im Wörterbuch vorhandenen Homographen durch eine Zählung abgegrenzt. Im Wörterbuch werden hierzu hochgestellte arabische Ziffern verwendet. Um deutlich zu machen, dass es sich nicht um Homographen handelt, die auch im Wörterbuch vorkommen, werden beim Neuansetzen griechische Buchstaben (Minuskeln:α, β, γ, δ, ε, ζ, η, θ usw.) verwendet.

Beispiel: kêsena ‘Käse herstellen’
vgl. das bereits im LBCM aufgeführte homographe kêsen ‘auswählen’

3 Regeln für spezielle Fälle

3.1 Wortartwechsel

Bei Fällen des Wortartwechsels wird stets die dem Wechsel zugrunde liegende Wortart lemmatisiert.

Token (Trans)	PoS-Tag	Lemma
to	...	...
etende	NA < VVINF	ēten

Table: Beispiel für Wortartwechsel NA < VVINF

oder

Token (Trans)	PoS-Tag	Lemma
irganghene	ADJA < VVPP	ergân

Table: Beispiel für Wortartwechsel ADJA < VVPP

3.2 Negierung

Bei negierten Adjektiven, die auf ein VVPP zurückgehen, wird die positive Infinitiv-Form lemmatisiert.

Token (Trans)	PoS-Tag	Lemma
ungewaschen	ADJD < VVPP	waschen

Table: Beispiel für Negierung

4 Hinweise zu einzelnen Wortarten

4.1 Verben

4.1.1 Grundsätzliches

Verben werden stets in ihrer **infinitivischen Form** lemmatisiert.

4.1.2 Zu Beachtendes

Bei konjugierten Verben oder Partizipien (sei es beim VVPS oder beim VVPP) wird stets die infinitivische Form lemmatisiert.

Bei Unsicherheiten in der Zuordnung von `ge-`Formen muss sich für einen Lemmaeintrag entschieden werden, z.B. kann das Token `ghescen` entweder das Lemma `schên` oder das Lemma `geschên1` erhalten.

4.1.3 Partikelverben

Bei getrennt stehenden, univerbierten (d.h. ins Wörterbuch eingetragenen) Partikelverben erhält das Verb das Lemma des Partikelverbs (z.B. `sach` erhält das Lemma `ansên`), die Partikel (z.B. `an`) hingegen wird nicht lemmatisiert(siehe auch Kap. Ausgeschlossene Token). Bei getrennt oder zusammen stehenden Partikelverben, die noch nicht univerbiert sind (d.h. nicht ins Wörterbuch eingetragen sind), werden das Verb und das Adverb getrennt voneinander lemmatisiert (z.B. `wech varn` als Adverb `wech` und Verb `vāren1`). Dies gilt auch für analog gebildete Konstruktionen.

4.2 Determinierer/ Pronomen

4.2.1 Grundsätzliches

Determinierer und Pronomen werden stets in ihrer **nominativischen Form** lemmatisiert.

4.2.2 Zu Beachtendes

4.2.3 Pronomen der ersten und zweiten Person

Personalpronomen der ersten und zweiten Person werden stets mit der **nominativischen Form im Singular** oder **im Plural** je nach Auftreten im Text lemmatisiert.

Flektierte Formen werden, auch wenn sie fettgedruckt im Wörterbuch erscheinen, als Variantenlemmata behandelt. Veröffentlicht werden ausschließlich die nominativischen Formen (z.B. `ik`, aber nicht `mî`).

Possessivpronomen der ersten und zweiten Person werden wie die Personalpronomen stets mit der **nominativischen Form im Singular** oder **im Plural** je nach Auftreten im Text lemmatisiert (z.B. `mîn2`, `dîn`, `iuwe`).

4.2.4 Determinierer und Pronomen mit Portmanteau-Formen

Der Definitartikel, der Indefinitartikel, die Personalpronomen der 3. Person Singular, die Demonstrativa sowie die Possessiva der 3. Person Singular erhalten Lemmata mit Portmanteau-Formen:

Bezeichnung: PoS-Tag	veröffentlichtes Lemma	Beispiel
Definitartikel: DDART...	dê1/dê1/dat2A	*in [deme] huse*
Indefinitartikel: DIART...	ê^{i}n1/ê^{i}ne1/ê^{i}n1	*in [ener] wise*
Personalpronomen der dritten Person: PPER	hê1/sê1/et1	*[he] seget*
Demonstrative Determinierer/ Pronomen: DD..., DPDS	desse/desse/dit	*[desser] lude*
Possessive Determinierer/ Pronomen der dritten Person Sg.: DPOS..., DPPOSS	sîn4/ēr1/sîn4	*to [sinen] broderen*

Table: Liste der Portmanteau-Formen

Personalpronomen der 3. Person mit einer Numerusambiguität (wenn z.B. bei der Form `en` trotz des Kontextes eine eindeutige Bestimmung als Singular `hê1` oder Plural `sê2` nicht möglich ist) erhalten das Lemma `hê1/sê1/et1/sê2`.

5 Anhang

5.1 B Grundlage für die digit. Lemmaliste

- Damme, Robert (2011): *Vocabularius Theutonicus*. Überlieferungsgeschichtliche Edition des mittelniederdeutsch-lateinischen Schulwörterbuchs. 3 Bde. (Niederdeutsche Studien, 54). Köln, Weimar, Wien: Böhlau.
- LBCM = Mittelniederdeutsches Handwörterbuch. Begr. v. Agathe Lasch u. Conrad Borchling. Fortgef. v. Gerhard Cordes u. Dieter Möhn. Neumünster 1928 ff.: Wachholtz.

- LW = Mittelniederdeutsches Handwörterbuch v. August Lübben. Nach dem Tode der Verf. vollend. v. Christoph Walther. Darmstadt 1995: Wiss. Buchges.

5.2 C Literatur

- Kleymann, Verena/ Nagel, Norbert/ Peters, Robert (2015): "Die digitale Lemmaliste für das Mittelniederdeutsche im DFG-Projekt 'Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650)'." In: Korrespondenzblatt des Vereins für niederdeutsche Sprachforschung (122,2), 95-100.