

NIEDERDEUTSCHES JAHRBUCH

Jahrbuch

des Vereins für niederdeutsche
Sprachforschung

Jahrgang 2017

140

WACHHOLTZ
MURMANN PUBLISHERS

Gedruckt mit Unterstützung
der Kulturbehörde Hamburg,
des Ministeriums für Justiz, Kultur und Europa
des Landes Schleswig-Holstein,
des Landschaftsverbandes Rheinland,
des Landschaftsverbandes Westfalen-Lippe



Für die Menschen.
Für Westfalen-Lippe.

Schriftleitung: Dr. Friedel Helga Rooffs
Kommission für
Mundart- und Namenforschung Westfalens
Schlossplatz 34
D-48143 Münster

ISSN 0083-5617
ISBN 978-3-529-04240-9

Alle Rechte, auch die des auszugsweisen Nachdrucks,
der fotomechanischen Wiedergabe und der Übersetzung, vorbehalten

Wachholtz Verlag – Murmann Publishers, Kiel/Hamburg
2017

INHALT

Niederdeutsche Etymologie: Neuere Forschungsansätze <i>Von Martin Kümmel</i>	7
Vom Nutzen des Altsächsischen für die Erforschung des altgermanischen Tempus- und Aspektsystems <i>Von Fabian Fleißner</i>	19
Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) <i>Von Robert Peters</i>	35
Historische Sprachdaten als Herausforderung für die manuelle und automatische Annotation: Das Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200–1650) <i>Von Ingrid Schröder, Fabian Barteld, Katharina Dreessen und Sarah Ihden</i>	43
Eine annotierte digitale Plattform für die Untersuchung der Stadtsprach- geschichte Greifswalds: Neue Methoden zur Erforschung des Niederdeutschen? <i>Von Matthias Schulz</i>	59
<i>Alle lude mane ik dar to, dat se dit bok nutten so, alse id en to den eren nicht missesta.</i> Untersuchungen zu einer verlorenen Sachsenspiegel- Handschrift <i>Von Jana Wolf</i>	79
Literatur der Altmark in Mittelalter und Früher Neuzeit. Eine Skizze <i>Von Volker Honemann</i>	95
Die „plattdeutschen Scribenten“ Groth, Reuter und Brinckman im literarischen Diskurs ihrer Zeit. Versuch einer multimodalen Perspektivierung <i>Von Barbara Scheuermann</i>	143

Buchbesprechungen

Timothy Sodmann (Hg.): Heliand. Der altsächsische Text/De Oudsaksi- sche tekst <i>Von Arend Quak</i>	177
--	-----

Gisbert Strottdrees: Im Anfang war die Woort. Flurnamen in Westfalen (Westfälische Beiträge zur niederdeutschen Philologie 16) <i>Von Michael Flöer</i>	178
Markus Denkler: Das münsterländische Platt (Westfälische Mundarten 1) <i>Von Alfred Lameli</i>	180
Kurt Wagner: Gorch Fock und Finkenwerder. Leben, Mythos, Vermächtnis (Sutton Geschichte) <i>Von Robert Langhanke</i>	181

Historische Sprachdaten als Herausforderung für die manuelle und automatische Annotation: Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)

*Ingrid Schröder, Fabian Barteld, Katharina Dreessen und Sarah Ihden,
Hamburg*

Die räumliche und zeitliche Ausgewogenheit des Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) sowie die Berücksichtigung des gesamten Textsortenspektrums ermöglichen es erstmals, sprachliche Varianz und Ambiguität im Mittelniederdeutschen auf breiter empirischer Basis systematisch zu erforschen. Dies ist für die Erarbeitung einer neuen mittelniederdeutschen Grammatik, die auf die diastematische Vielfalt sowohl in diatopischer wie diachroner und auch funktionaler Hinsicht einzugehen hat, eine unabdingbare Voraussetzung.¹

Varianz und Ambiguität des Mittelniederdeutschen, aber auch der mittelalterliche und frühneuzeitliche Schreibusus stellen spezifische Anforderungen an die korpuslinguistische Aufbereitung der Sprachdaten. Darauf sind Verfahren, die bisher für die Erstellung von Korpora moderner standardisierter Sprachen entwickelt worden sind, nicht ausgerichtet und erscheinen für die Anwendung auf historische Sprachen wenig geeignet. Mit Blick auf die spezifischen Bedarfe der Datenaufbereitung sind daher für das „Referenzkorpus Mittelniederdeutsch/Niederrheinisch“ (ReN) besondere Regeln aufgestellt und Werkzeuge entwickelt worden.

Bereits basale Vorverarbeitungsschritte, wie die Festlegung lexikalischer Einheiten (Tokenisierung) und die Bestimmung der Satzgrenzen, erweisen sich in den nicht-standardisierten Texten als nicht trivial und müssen im Zuge der Transkription als Vorbereitung auf die Annotation manuell erfolgen. Für die Annotation selbst ist ein spezifisches Tagset zu entwerfen, das einerseits möglichst genau den Befund charakterisiert, andererseits aber auch individuelle Interpretationen und somit Festlegungen so weit wie möglich vermeidet, die nicht aus der Form des sprachlichen Materials eindeutig ablesbar und damit intersubjektiv überprüfbar sind, sondern auf (ggf. nicht zutreffenden) linguistischen Vorannahmen der AnnotatorInnen beruhen

1 Zum Korpus als grammatikographische Basis vgl. Schröder 2014.

(z. B. zur Kasusforderung von Präpositionen). Besonders ist dies bei der Annotation der Wortarten (PoS-Annotation) zu beachten, da u. a. eine Unterscheidung von Konjunktionen und Subjunktionen und damit zugleich von Haupt- und Nebensätzen durch die im Vergleich zum Neuhochdeutschen freiere Verbstellung häufig nicht eindeutig zu treffen ist. Die morphologische Annotation wird durch ambige Formen (z. B. Genus-, Kasus- und Numerusambiguität) erschwert, wenn der sprachliche Kontext keine Disambiguierung ermöglicht. Auch gängige Prüfverfahren, wie z. B. Umstellproben, die eine muttersprachliche Kompetenz voraussetzen, können bei der Annotation nicht angewandt werden.

Fehlende Ressourcen erfordern zusätzliche Arbeiten zur Unterstützung des Annotationsprozesses. So wurde eine digitale Lemmaliste erstellt, welche im Annotationsprozess laufend erweitert wird. Zudem mussten vorhandene Annotationstools für das Mittelniederdeutsche angepasst und weiterentwickelt werden, da keine fertig nutzbaren Programme zur automatischen Annotation existieren.

Im Folgenden soll ein Überblick über die Phasen der Korpuserstellung im ReN den Workflow veranschaulichen. In weiteren Schritten werden die Lösungen vorgestellt, die im Projekt entwickelt worden sind, um den Herausforderungen einer nicht-standardisierten historischen Sprache, wie sie das Mittelniederdeutsche darstellt, begegnen zu können. Die Resultate der Datenaufbereitung und die Möglichkeiten für die Nutzung der annotierten Texte lassen sich abschließend durch einen Blick auf die verwendete Such- und Visualisierungssoftware „ANNIS“ (Krause/Zeldes 2016) veranschaulichen.²

1 Korpuserstellung

Bei der Korpuserstellung durchläuft jeder Text zwei Phasen, nämlich die Transkription, mit der die Textaufbereitung verbunden ist, und die Annotation, bevor sich nach der Publikation die Nutzungsphase anschließt. Als erste Stufe der Textaufbereitung erfolgt im Zuge der Transkription die Tokenisierung, d. h. die Festlegung von Worteinheiten, und die Satzgrenzenfestlegung. Zur Annotation gehören die Zuweisung der syntaktischen Wortart³ durch das „Part of Speech“ (PoS)-Tagging, die Eingabe von Informationen zur Flexionsmorphologie und die Lemmatisierung⁴. In der dritten und letzten Arbeitsphase werden die Daten zum einen in „ANNIS“ (Krause/Zeldes

2 Für die Entwicklung der korpuslinguistischen Lösungsvorschläge ist das Teilprojekt Hamburg verantwortlich. Als zielführend hat sich innerhalb des Projekts die enge Bezugnahme computerlinguistischer und sprachhistorischer Expertise erwiesen. Die enge Zusammenarbeit mit dem Team des „Referenzkorpus Frühneuhochdeutsch“ ermöglichte es, auf dessen Entwicklungen aufzubauen. Dafür gilt allen Bochumer Kolleginnen und Kollegen großer Dank!

3 Siehe die Unterscheidung von lemma- und belegspezifischer Annotation bei Dipper u. a. 2013.

4 Auf die spezifischen Bedingungen der manuellen sowie automatischen Annotation bei nicht-standardisierten Sprachen wird in diesem Beitrag nicht eingegangen; vgl. dazu Barteld/Schröder/Zinsmeister 2016a; 2015; Barteld u. a. 2014.

2016), einer webbasierten Such- und Visualisierungssoftware, und zum anderen im „TEI“-Format (Text Encoding Initiative o. J.), einem Dokumentenformat zur Kodierung und zum Austausch von Texten, veröffentlicht.

Ein wesentliches Instrument für die Annotation der historischen Referenzkorpora stellt das webbasierte Annotationstool „CorA“ (Bollmann u. a. 2014) dar, das in Bochum für das „Referenzkorpus Frühneuhochdeutsch“ entwickelt wurde. Nach erfolgter Transkription⁵ werden die Texte in dieses Tool übertragen, das für die folgende Annotation eine Eingabemaske bietet und während des Annotationsprozesses auch eine Korrektur der Transkription und Tokenisierung erlaubt.

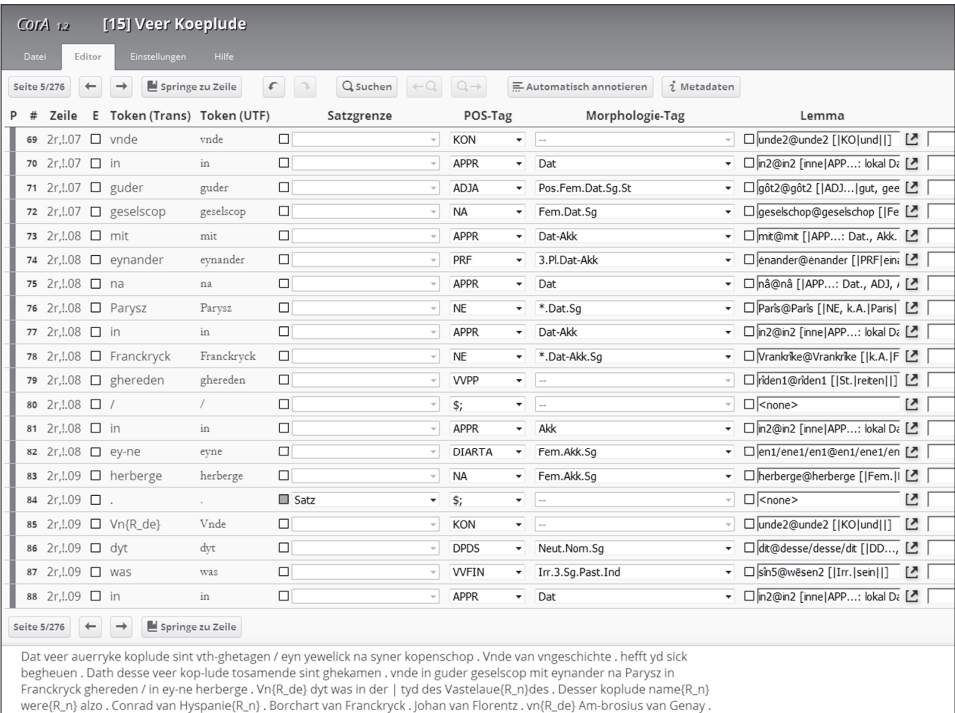


Abb. 1: Annotation mittels „CorA“ (Bildschirmausschnitt)

Von zentraler Bedeutung für die Annotation ist ein adäquates Tagset, das den spezifischen Erfordernissen genügt: Erstens soll es eine möglichst große Nähe zu existierenden Tagsets aufweisen, damit zu einem späteren Zeitpunkt eine gemeinsame Analyse der verschiedenen Referenzkorpora ermöglicht wird; zweitens müssen die Tags so definiert sein, dass sie durch den konkreten Kontext determiniert (und nicht etwa durch Intuition auszuwählen) sind; daraus ergibt sich, dass sie drittens die Ab-

5 Für die formalen Auszeichnungen, z. B. von Sonderzeichen, Abkürzungen oder Glossierungen, aber auch Seiten- und Zeilenwechseln, der Existenz von Abbildungen u. a. m., sind eigens Auszeichnungsroutinen entwickelt worden, die eine weitere elektronische Verarbeitung ermöglichen; für die Transkriptionsrichtlinien vgl. ReN-Team 2014.

bildung von Ambiguitäten ermöglichen müssen, aber viertens trotzdem so spezifisch wie möglich anwendbar sein sollen.

Dem Tagset für das ReN liegt das ebenfalls in Bochum für historische Belange entwickelte „Historische Tagset“ (HiTS; vgl. Dipper u. a. 2013) zugrunde, das wiederum auf dem „Stuttgart-Tübingen Tagset“ (STTS; vgl. Schiller u. a. 1999) beruht. Durch Anpassung an die Spezifika des Mittelniederdeutschen ist das „Historische Niederdeutsch-Tagset“ (HiNTS; vgl. ReN-Team 2016a) entstanden. Es bietet u. a. den Vorteil einer stark kontextbasierten Annotation und dient so der Vermeidung (möglicherweise falscher) Vorannahmen und Interpretationen. In der Klasse der indefiniten Determinierer und Pronomen bspw. wird nicht vorab festgelegt, welche Lexeme ausschließlich substituierend und somit als Pronomen im eigentlichen Wortsinn auftreten, da dies erst im Ergebnis der Korpusauswertung entscheidbar ist.

Für den Erfolg des Projekts sind klare Richtlinien für Transkription und Annotation entscheidend, um die Homogenität der Daten sicherzustellen. Diese Richtlinien sind detailliert in einem Transkriptions- und zwei Annotationshandbüchern (PoS und Morphologie; Lemmatisierung) dokumentiert (vgl. ReN-Team 2014; 2016a; 2016b). Die Annotationsrichtlinien wurden mehrfach mithilfe von „Inter-Annotator Agreement“-Experimenten, in denen die Abweichungen zwischen den Entscheidungen verschiedener AnnotatorInnen gemessen werden, überprüft und entsprechend angepasst.⁶ Die ersten Experimente im ReN-Projekt deckten als die häufigste Ursache für Differenzen zwischen den AnnotatorInnen fehlende Richtlinien für spezifische Fälle auf. Indem das Handbuch um entsprechende Regelungen ergänzt wurde, konnte diese Ursache für Abweichungen in den Annotationen deutlich reduziert und somit die Qualität der Annotation sichergestellt werden. Das Ergebnis ist außerdem eine hohe Konsistenz der Korpusauszeichnung. Transparenz ist ein zentraler und unabdingbarer Grundsatz der Arbeit, um den NutzerInnen des Korpus die Möglichkeit zu eröffnen, alle Entscheidungen nachzuvollziehen und zu überprüfen.

Die manuelle Annotation wird durch automatisch erstellte Annotationsvorschläge unterstützt. Hierfür werden statistische Annotationstools eingesetzt. Da vor allem die graphische Variation im Mittelniederdeutschen den Einsatz von Annotationstools erschwert, die für standardisierte Sprachen entwickelt worden sind, sind auch diese Tools an die spezifischen Erfordernisse bei der Annotation der mittelniederdeutschen Texte angepasst worden.⁷ Die Annotationstools werden vor ihrem Einsatz auf per Hand annotierten Daten trainiert und in regelmäßigen Abständen auf den nunmehr automatisch annotierten und manuell korrigierten Daten neu trainiert, so dass sich die Qualität der Vorschläge sukzessiv verbessert (vgl. Abb. 2).

6 Inter-Annotator Agreement unter Anwendung des MAMA-Zyklus (Model-Annotate-Model-Annotate, vgl. Pustejovsky/Stubbs 2012: 27, 109). Zur Weiterentwicklung im Projektzusammenhang vgl. Barteld/Schröder/Zinsmeister 2016b.

7 Für die grammatische Annotation wird der „RFTagger“ (Schmid/Laws 2008) genutzt, der Lemmatisierung liegt das Tool „Lemming“ zugrunde (Müller u. a. 2015); zur Anpassung an das Mittelniederdeutsche vgl. Barteld/Schröder/Zinsmeister 2015; 2016a.

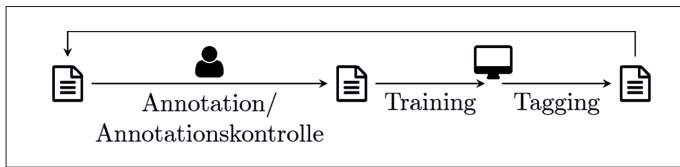


Abb. 2: Annotations-Workflow

2 Textaufbereitung

Im Rahmen der Textaufbereitung erfolgen, wie bereits erwähnt, die Tokenisierung sowie die Satzgrenzenfestlegung. Die Tokenisierung ist zwingend erforderlich, da die festgelegten Worteinheiten die Grundlage für die spätere grammatische Annotation darstellen. Anders als die Tokenisierung ist die Satzgrenzenfestlegung zwar nicht obligatorisch, stellt aber sowohl für die manuelle als auch für die automatische Annotation eine erhebliche Unterstützung dar.

Die Entscheidung, was eine Worteinheit umfasst, ist in historischen Texten keineswegs trivial. Bereits die Identifizierung von Wortformen wird dadurch behindert, dass einerseits die Einheiten nicht konsequent durch Spatien voneinander getrennt werden, andererseits Spatien nicht auf mehrere Wortformen hindeuten müssen, beispielsweise bei getrennt geschriebenen Komposita. Somit ist zwischen graphischen Einheiten im Originaltext und Wortformen als Annotationsgrößen konsequent zu unterscheiden.

Als Ausgangspunkt für die Identifizierung von Wortformen dient die Lemmatisierung im „Mittelniederdeutschen Handwörterbuch“ (Lasch/Borchling 1956ff.). Daher wird z. B. die klitisierte Negationspartikel *en* als eine eigenständige Einheit erfasst (vgl. ebd., s. v.), eine Entscheidung, die auch anders ausfallen könnte, für die Korpuszwecke aber angemessen erscheint.

Am Beispiel des „Oldenburger Sachsenspiegels“ (vgl. Abb. 3)⁸ soll das Vorgehen erläutert werden. In allen Fällen von getrennt geschriebenen Elementen eines Lexems wie bei den Präfigierungen in *be scede* und *ghe don* oder auch bei den beiden Morphemen in *al ene* sind diese jeweils zu Worteinheiten zusammenzufassen, damit sie als Ganzes annotiert werden können. Zugleich muss aber auch die Information der Getrenntschreibung erhalten bleiben, damit eine an der Originalschreibung orientierte Textfassung publiziert werden kann. Im Transkript signalisiert in solchen Fällen ein Raute-Zeichen, dass es sich um eine Auseinanderschreibung im Original handelt, die zum Zwecke der Annotation zusammengezogen wird, z. B. *be#scede*, *ghe# don*, *al# ene*. Dasselbe gilt für die getrennt geschriebenen Bestandteile von Komposita. Da Kandidaten für Komposita im Mittelniederdeutschen nicht immer eindeutig in ihrem Status als Wortform bestimmt werden können, sind zu-

8 Landesbibliothek Oldenburg, Cim I 410, Bl. 6r, Leihgabe der Niedersächsischen Sparkassenstiftung; Abb. mit freundlicher Genehmigung der Niedersächsischen Sparkassenstiftung.

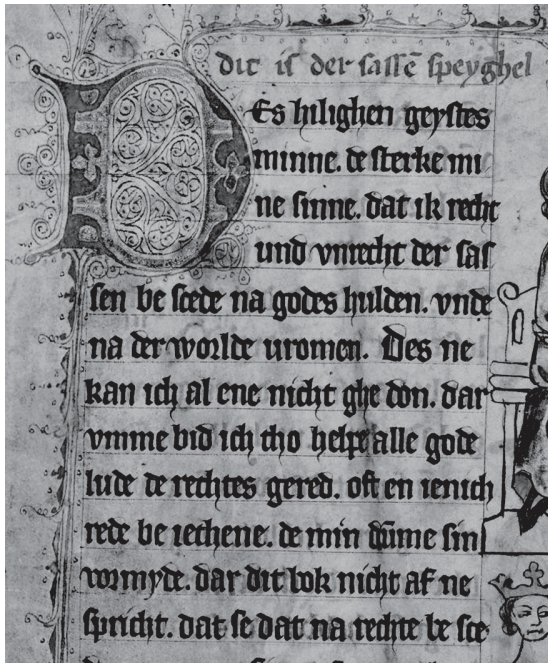


Abb. 3: „Oldenburger Sachsenspiegel“, Bl. 6r

sätzlich auch der Kontext und die grammatischen Merkmale für eine Entscheidung heranzuziehen.

Sind Elemente zusammengeschrieben, die als zwei getrennte lexikalische Einheiten zu bestimmen sind, sowohl bei fehlenden Spatien zwischen zwei Wortformen als auch bei Klitisierungen, werden die einzelnen Bestandteile getrennt, indem ein Paragraphenzeichen eingefügt wird (z. B. *enſbrachte*, *sprekstſu*). Daraus generiert das Annotationstool später automatisch einzelne Wortformen, die dann gesondert annotiert werden können. In der in „ANNIS“ publizierte Version des Textes sind dann beide Varianten anzeigbar, d. h. die Trennung sowohl nach der originalen Schreibung als auch nach Wortformen.

Für die linguistische Größe „Satz“ spiegelt der Eintrag im „Lexikon der Sprachwissenschaft“ die Uneindeutigkeit der Bestimmung wider: „Nach sprachspezifischen Regeln aus kleineren Einheiten konstruierte Redeeinheit, die hinsichtlich Inhalt, gramm. Struktur und Intonation relativ vollständig und unabhängig ist.“ (Bußmann 2008: 601) Ein zuverlässiges, wenn auch nicht immer ausreichendes Kriterium ist die Existenz eines finiten Verbs, wie es die IDS-Grammatik in ihrer Satzdefinition zum Ausdruck bringt: „Sätze enthalten ein finites Verb und (in der Regel) die unter strukturellen und kontextuellen Bedingungen notwendigen Komplemente dieses Verbs; darüber hinaus können Sätze Supplemente enthalten.“ (Zifonun u. a. 1997: 87) Hieran anschließend, gilt auch im ReN-Projekt als Satz jede grammatische Einheit, die ein finites Verb enthält. Dabei wird in Kauf genommen, dass satzförmige Einheiten ohne finites Verb auf der Annotationsebene nicht als Satz ausgewiesen

werden, so z.B. Überschriften, die daher gesondert als solche gekennzeichnet werden.

Ob es sich bei einer Satzeinheit um einen Haupt- oder Nebensatz handelt, kann im Mittelniederdeutschen aufgrund der freieren Verbstellung nicht immer eindeutig entschieden werden (vgl. Hård 2000: 1461),⁹ so dass auf eine formale Differenzierung von Haupt- und Nebensätzen verzichtet werden muss. Alle Satzgrenzen werden, unabhängig davon, ob im Originaltext ein Satzzeichen vorhanden ist (das ohne weitere Kennzeichnung transkribiert wird), einheitlich durch „\$.“ gekennzeichnet, z.B. *DEs hilighen geystes minne . de sterke mi# ne sinne . \$. dat ik recht und vnrecht der sas# sen be# scede na godes hulden . vnde na der worlde uromen . \$. Des ne kan ich al# ene nicht ghe# don . \$.* („Oldenburger Sachsenspiegel“, Bl. 6r, vgl. Abb. 3). Satzeinheiten, die infolge von Einschüben getrennt sind, müssen markiert werden. Dies geschieht durch „\$.a\$“, „\$.b\$“ usw., z.B. *Vnde alle \$.a\$ de in deme sale weren \$. vorschreckeden . \$.b\$* („Veer Koeplude“, Bl. 7r).¹⁰

Erst diese Maßnahmen gewährleisten, dass der Annotation diskrete Elemente in einheitlicher Weise zugrunde liegen, welche eine Grundbedingung für die weitere korpuslinguistische Aufbereitung darstellen. Dadurch werden morphologisch und syntaktisch beschreibbare Einheiten unterschieden, auf die mit automatischer Unterstützung zugegriffen werden kann.

3 Annotation

Im Bereich der Annotation führen die Besonderheiten der nicht-standardisierten historischen Texte zu Zweifelsfällen sowohl hinsichtlich der Feststellung der Wortart (PoS-Annotation) als auch hinsichtlich der Bestimmung der Flexionsmorphologie (morphologische Annotation). Zweifel bei der Feststellung der Wortart treten z.B. bei der Unterscheidung von Konjunktion und Subjunktion auf, da bei der Identifizierung von Haupt- und Nebensätzen gängige Verfahren wie die Analyse der Wortstellung nicht ohne Einschränkungen greifen (vgl. oben). Zudem kann dasselbe Lexem sowohl Hauptsätze wie Nebensätze einleiten, wie es das „Mittelniederdeutsche Wörterbuch“ von Schiller/Lübben beispielsweise für *wente* ausweist, das sowohl ‘denn’ als auch ‘weil’ bedeuten kann, aber auch ‘dass’ und ‘aber’ (vgl. Schiller/Lübben 1875–80, s. v.). Diese Ambiguität wird auch in grammatischen Darstellungen zum Mittelniederdeutschen dokumentiert:

Dem durch das begründende *wente* eingeleiteten Satz geht in der Regel der begründete Satz voraus, und *wente* hat koordinierende Funktion; in vielen Texten

9 Hård (2000: 1461) konstatiert, dass es „oft keine strenge syntaktische Unterscheidung (durch Wortfolge) [...] zwischen Hauptsatz und Nebensatz“ gibt.

10 Diese Auszeichnung der Satzgrenzen wird für die Annotation mittels „CorA“ in Satzgrenzen-Markierungen, die am letzten Token eines Satzes gesetzt werden, umgewandelt (vgl. Abb. 1). In der Publikation in „ANNIS“ werden die Sätze als Spannen dargestellt, die sich jeweils über die einzelnen Satzeinheiten erstrecken.

leitet *wente* aber einen subordinierten Satz ein, der dann dem übergeordneten, begründeten, vorangehen kann. (Härd 2000: 1461)

Je nach Stellung des Teilsatzes kann *wente* sowohl koordinierende als auch subordinierende Funktion haben, wobei nicht ausgeschlossen ist, dass ein subordinierter *wente*-Satz auch nachgestellt ist.

Zur Veranschaulichung des Vorgehens dient ein Beispiel aus dem „Johannes-evangelium“ des „Buxtehuder Evangeliars“: *vnde ik sach et · vnde betugede et · wente dit is godes sone* („Buxtehuder Evangeliar“, S. 6). Da es sich hierbei um einen nebengeordneten oder untergeordneten Kausalsatz handeln könnte, müssen für die Annotation rein strukturelle Regeln zugrunde gelegt werden, um eine konsequente Annotation sicherzustellen, die unabhängig von der jeweiligen individuellen Interpretation der AnnotatorInnen bleibt. Im Fall der Junktionen beinhalten die Annotationsregeln (ReN-Team 2016a), dass ein als Konjunktion (KON) annotiertes Element im Vorvorfeld steht oder gleichrangige Elemente wie Nominalphrasen (z. B. *man unde wip*) verbindet, eine Annotation als Subjunktion (KOUS) hingegen bei Verbspät- oder Verbletzstellung erfolgt. Die Entscheidung fällt in dem oben genannten Fall demnach für die nebenordnende Konjunktion.

Die Kenntnis dieser Konvention ist entscheidend, wenn später mithilfe des Korpus grammatische Analysen vorgenommen werden. Denn dann ist zu berücksichtigen, dass KON nicht etwa bedeutet, dass tatsächlich ein nebengeordneter Satz vorliegt, sondern lediglich, dass die Junktion im Vorvorfeld steht und/oder mit V2-Stellung verbunden ist. Die Frage, ob im Mittelniederdeutschen solche Fälle auch als Nebensätze zu interpretieren sind, kann und soll bei der Korpuserstellung nicht beantwortet werden. In Sätzen, die nur aus Subjekt und Prädikat bestehen und daher topologisch ambig sind, wie in **wente he sleep*, wird das Portmanteau-Tag „KO*“ benutzt, das die Entscheidung offen lässt.

Ein zweites Beispiel, das gleichzeitig Unterschiede zwischen „HiTS“ und „HiNTS“ verdeutlicht, betrifft die Klasse der Determinierer und Pronomen. Ebenso wie in „HiTS“ werden auch bei „HiNTS“ die Tags durch ein „A“ für „attribuierend“ oder ein „S“ für „substituierend“ spezifiziert. Allerdings werden bspw. in der Gruppe der indefiniten Determinierer und Pronomen im ReN-Projekt lexembezogene Vorannahmen vermieden. Statt vorab festzulegen, welche Lexeme sowohl attribuierend als auch substituierend und welche ausschließlich substituierend vorkommen können, wird in „HiNTS“ allein aufgrund des gegebenen Kontextes entweder „attribuierend“ oder „substituierend“ annotiert. Auf diese Weise können in späteren Analysen diejenigen Lexeme ermittelt werden, die ausschließlich substituierend vorkommen und somit als Pronomen im eigentlichen Wortsinn zu klassifizieren sind.

Im Bereich der flexionsmorphologischen Annotation sind häufiger ambige Formen anzutreffen, z. B. beim Genus. In der Überschrift *Dit is der sassen speyghel* („Oldenburger Sachsenspiegel“, Bl. 6r, vgl. Abb. 3) betrifft dies das Substantiv *speghel*, das laut Wörterbucheintrag sowohl Maskulinum als auch Neutrum sein kann (vgl. Lasch/Borchling 1956ff., s. v.). In Fällen wie diesem, in denen der sprachliche Kontext keine Disambiguierung ermöglicht, wird mithilfe von Portmanteau-Tags (hier: Masc-Neut) die Ambiguität abgebildet.

Eine weitere Schwierigkeit ergibt sich bei Ambiguitäten mit Korrelation von Genus, Kasus und Numerus. Im Satz *den wy vmme armoth siner olderer van barmherticheyt weghen van enen kleuen (!) kinde hebben vpgetagen* („Griseldis“, Bl. 19r) kann das Substantiv *armôt* laut „Mittelniederdeutschem Handwörterbuch“ (Lasch/Borchling 1956ff., s. v.) Maskulinum oder Femininum sein. Auch diese Ambiguität kann nicht mithilfe des sprachlichen Kontextes aufgelöst werden. Handelt es sich um ein Maskulinum, kann die Form für den Akkusativ Singular stehen. Handelt es sich jedoch um ein Femininum, kann *armoth* den Genitiv, Dativ und auch Akkusativ Singular darstellen. Aufgrund der Präposition *umme* (ebd., s. v.), die den Dativ oder den Akkusativ regiert, kann der Synkretismus zumindest auf diese beiden Kasus reduziert werden, die beide im Tag (Dat-Akk) zum Ausdruck gebracht werden. Da die vorhandene Korrelation zwischen Genus und Kasus nicht mithilfe der Tags abgebildet werden kann, wird als zusätzliche Information in die Kommentarspalte „Fem. Dat-Akk. oder Masc.Akk.“ geschrieben.

Die Annotation zielt also wesentlich darauf ab, alle Zuordnungen zu Wortarten und alle Informationen zu den morphologischen Ausprägungen möglichst vollständig und exakt zugänglich zu machen. Dafür werden klare und eindeutige Annotationsregeln aufgestellt, die gleichzeitig verhindern, dass Zuordnungen intuitiv vorgenommen werden, die nicht durch die Form gestützt sind. Auf diese Weise kann gewährleistet werden, dass späteren grammatischen Analysen nicht unzulässig vorgegriffen wird.

4 Publikation und Nutzung

In der dritten und letzten Phase werden die Daten zur Veröffentlichung in ein von „ANNIS“, der webbasierten Such- und Visualisierungssoftware, lesbares Format und in ein „TEI“-Format überführt. Um die Nutzbarkeit der Daten möglichst früh zu gewährleisten, werden Vorabversionen des Korpus veröffentlicht. Zum Zeitpunkt der Bearbeitung dieses Aufsatzes (April 2017) sind zwei Versionen (2016-08-23 und 2017-03-08) mit 5 bzw. 18 Texten veröffentlicht. Die Vorabversionen bleiben auch nach der Publikation der finalen Version des Korpus verfügbar und können daher bereits für reproduzierbare Untersuchungen genutzt werden. Mit Hilfe von „ANNIS“ können sowohl das Gesamtkorpus als auch Teilkorpora oder einzelne Texte nach allen Elementen der Transkriptionsebene wie auch der Annotationsebenen durchsucht werden. Zu einem späteren Zeitpunkt wird auch eine Abfrage in mehreren Referenzkorpora zugleich möglich sein, so dass sprachliche Phänomene verschiedener Schreibregionen oder Zeitstufen miteinander verglichen werden können. Als Beispiele für die Abfragemöglichkeiten dienen im Folgenden eine Wortartensuche und eine Suchabfrage nach der Kasusrektion.¹¹

Will man beispielsweise in Erfahrung bringen, wie häufig eine Wortart im Gesamtkorpus oder in einem Text auftritt (im Beispiel Eigennamen „NE“ im Text „Griseldis“ in der Korpusversion "ReN_2017-03-08"), lässt sich eine entsprechende

11 Zur Benutzungsanleitung für „ANNIS“ vgl. Zeldes 2016.

Suchabfrage in „ANNIS“ formulieren (vgl. Abb. 4). In der Ein- und Ausgabemaske befindet sich oben links das Feld für die Suchabfrage in formalisierter Form, rechts die Ergebnisausgabe. Die Ergebnisse werden im KWIC [keyword-in-context]-Format ausgegeben und lassen sich zusätzlich quantitativ aufbereiten, so dass die Frequenz der ermittelten Elemente in den einzelnen Texten untersucht werden kann.

The screenshot shows the ANNIS web interface. On the left, the 'Query Builder' section contains a search query: `pos="NE" & meta::doc="Griseldis"`. Below the query builder, it shows '89 matches in 1 document'. A table lists the results, with the first row highlighted: 'ReN_2017-03-08' with 18 texts and 121,357 tokens. On the right, the 'Query Result' section displays the search results in a KWIC format. It shows four results, each with a path and a keyword. The first result is 'Path: ReN_2017-03-08 > Griseldis (tok_dipl 5 - 15)' with the keyword 'der'. The second result is 'Path: ReN_2017-03-08 > Griseldis (tok_dipl 24 - 34)' with the keyword 'Vnde'. The third result is 'Path: ReN_2017-03-08 > Griseldis (tok_dipl 28 - 38)' with the keyword 'wen'. The fourth result is 'Path: ReN_2017-03-08 > Griseldis (tok_dipl 29 - 39)' with the keyword 'Griseldis'. Each result is followed by a list of annotations: Transcription, Lexical Annotation, Graphical Annotation, Sentence Annotation, and All.

Abb. 4: „ANNIS“: Suchabfrage zu Elementen einer Wortart (Bildschirmausschnitt)

Eine solche quantitative Aufbereitung veranschaulicht die Abb. 5 anhand der Suche, mit welchen Kasus die Präposition *in* im Text „Griseldis“ steht. Das Ergebnis ist in Form eines Säulendiagramms und einer Tabelle wiedergegeben. Es weist aus, dass der Dativ klar mit 87 Fällen überwiegt, in 49 Fällen der Akkusativ der Präposition folgt und es in 34 Fällen formal nicht entscheidbar ist, ob es sich um einen Akkusativ oder um einen Dativ handelt.

Sollen die Belege für *in* mit Akkusativ näher betrachtet werden, so können diese als Liste ausgegeben werden (vgl. Abb. 6). Die Präposition steht in der ersten Spalte, danach folgt die Nominalphrase mit Artikel und Nomen. Die einzelnen Zeilen zeigen die möglichen Darstellungsformen der Annotation: 1. Zeile (tok_anno¹²): Wortformen als bearbeitetes Token (d.h. nach erfolgter Wortgrenzenfestlegung); 2. Zeile (lemma_wsd): Lemmata mit Homonymendisambiguierung gemäß dem entsprechenden Wörterbucheintrag; 3. Zeile (lemma): Lemmata ohne Homonymenkennzeichnung in der Schreibweise des „Mittelniederdeutschen Handwörterbuchs“

12 In der ersten Korpusversion (2016-08-23) heißt diese Ebene „tok_mod“.

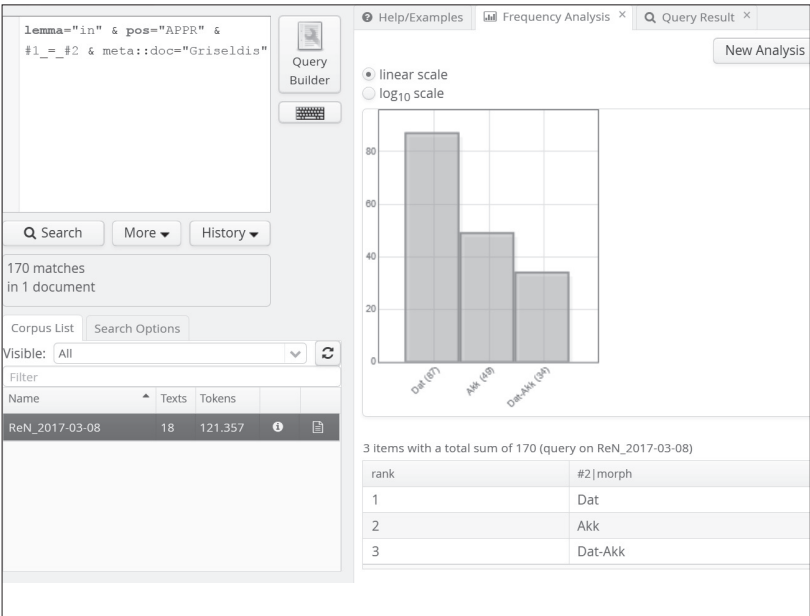


Abb. 5: „ANNIS“: Suchabfrage zu Kasus nach *in* (Bildschirmausschnitt)

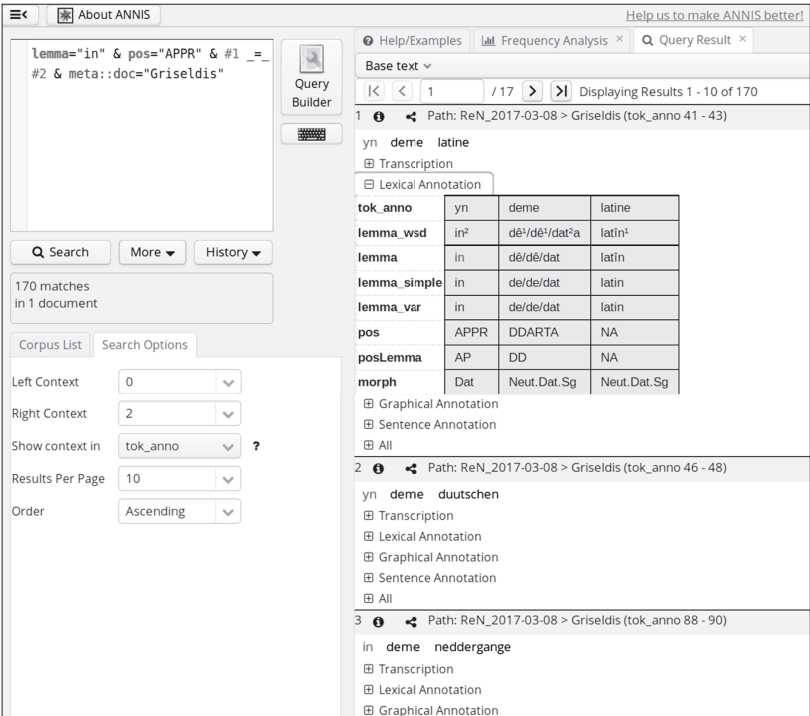


Abb. 6: „ANNIS“: Ergebnisdarstellung in Listenform (Bildschirmausschnitt)

inklusive der Längenauszeichnungen; 4. Zeile (lemma_simple): Lemmata ohne Diakritika; 5. Zeile (pos): syntaktische Wortarten (APPR = Präposition; DDARTA = Determinativ, definit, artikelartig, vorangestellt; NA = Nomen appellativum); 6. Zeile (posLemma): Wortarten (AP = Adposition; DD = Determinativ, definit; NA = Nomen appellativum); 7. Zeile (morph): morphologische Annotationen (Dat = Dativ; Neut.Dat.Sg = Neutrum Dativ Singular). Neben diesen Ausgabemöglichkeiten der lexikalischen Annotation (Lexical Annotation) erlaubt „ANNIS“ auch die Ausgabe des Transkripts (Transcription) oder anderer Annotationsebenen (Graphical Annotation, Sentence Annotation).

5 Erträge

Mit dem ReN wird der historischen Sprachforschung ein Korpus diplomatisch transkribierter, lemmatisierter und grammatisch annotierter mittelniederdeutscher und niederrheinischer Texte digital zur Verfügung gestellt. Mit seiner Hilfe werden vor allem grammatische Analysen ermöglicht, die bisher mangels geeigneter Datengrundlage nur ausschnittshaft realisiert werden konnten. Vorzüge des Korpus als Analysegrundlage bestehen in der konsistenten Datenbasis mit einer gleichmäßigen Verteilung in Raum und Zeit und unter Berücksichtigung der relevanten Gebrauchsdomänen, die im Korpus als „Felder der Schriftlichkeit“ ausgewiesen sind.

Die reflektierte Annotationsweise lässt es zu, diatopische, diachrone und diasituative strukturelle Differenzen herauszufiltern, die der Analyse zugänglich gemacht werden, ohne dass sie durch Präjudiz der Editoren verschleiert sind. Dafür ist vor allem durch die transparente Darstellung mit der detaillierten Dokumentation in den Transkriptions- und Annotationshandbüchern gesorgt, die jede Annotationsentscheidung nachvollziehbar macht.

Das Referenzkorpus füllt auf diese Weise auch eine wesentliche Lücke hinsichtlich der Formulierung von Transkriptions- und vor allem Annotationsstandards in Bezug auf das Mittelniederdeutsche. Hinzu kommen Programm(weiter)entwicklungen, u. a. zum Inter-Annotator Agreement und zur Lemmatisierung, die speziell auf die Bedarfe einer nicht-standardisierten historischen Sprachstufe zugeschnitten sind.¹³

Ein besonderer Vorzug liegt in der Übertragbarkeit der Methoden auf andere nicht-standardisierte Sprachformen wie beispielsweise mündliche Äußerungen oder auch wenig regulierte Sprachprodukte wie Chat-Foren, in denen Merkmale der konzeptionellen Mündlichkeit und Schriftlichkeit kombiniert werden.

13 Zusätzlich zu den in diesem Beitrag vorgestellten Methoden wurden auch weitere Tools entwickelt und angepasst. So wurde an einem Programm zur Überführung der Daten von „CorA“ in „ANNIS“ mitgearbeitet („Pepper-Modul“ für „CorA-XML“, <https://github.com/korpling/CorAXMLModules>) sowie eine Anpassung des Datenexports aus „ANNIS“ an die im Projekt verwendete doppelte Tokenisierung vorgenommen.

Quellen

- Buxtehuder Evangeliar = Qvatuor Evangeliorum versio Saxonica. [Hs. um 1480.]
Königliche Bibliothek Kopenhagen: Thott 8, 8°.
- Griseldis = De historie van der duldicheyt der vrowen geheten Griseldis; Uan sigismunda Des vorsten dochter van Salernia Unde van deme iungelinge Gwiscardo.
Hamburg: Drucker des Jegher 1502. Königliche Bibliothek Kopenhagen: 77¹-12, 4°.
- Oldenburger Sachsenspiegel = Oldenburger Bilderhandschrift des Sachsenspiegels.
Rastede 1336. Landesbibliothek Oldenburg: CIM I 410.
- Veer Koeplude = Historie van veer Koepluden vnde eyner thuchtigen vramen Vrouwen.
Hamburg: Borchard 1510. Staats- und Universitätsbibliothek Hamburg: Scrin A/226.
- Für weitere Informationen s. Metadaten im Korpus, Release 2016-08-23 (URL: <http://hdl.handle.net/11022/0000-0001-B002-5>).

Literatur

- Barteld, Fabian, Ingrid Schröder und Heike Zinsmeister (2015): Unsupervised regularization of historical texts for POS tagging. In: Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH). 10 December 2015, Warsaw, Poland, S. 3–12. URL: http://crh4.ipipan.waw.pl/files/9814/4973/5451/CRH4_proceedings.pdf [Stand: 23.03.2017].
- Barteld, Fabian, Ingrid Schröder und Heike Zinsmeister (2016a): Dealing with word-internal modification and spelling variation in data-driven lemmatization. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016). August 11, 2016, Berlin, Germany, S. 52–62. URL: <http://anthology.aclweb.org/W16-2106> [Stand: 23.03.2017].
- Barteld, Fabian, Ingrid Schröder und Heike Zinsmeister (2016b): text-gamma – Inter-annotator agreement for categorization with simultaneous segmentation and transcription-error correction. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). September 19–21, 2016, Bochum, Germany, S. 27–37. URL: https://www.linguistics.rub.de/konvens16/pub/4_konvensproc.pdf [Stand: 23.03.2017].
- Barteld, Fabian u. a. (2014): Annotating descriptively incomplete language phenomena. In: LAW VIII – The 8th Linguistic Annotation Workshop in conjunction with COLING 2014. Proceedings of the Workshop. August 23–24, 2014, Dublin, Ireland, S. 99–104. URL: <http://www.aclweb.org/anthology/W14-4915> [Stand: 23.03.2017].
- Bollmann, Marcel u. a. (2014): CorA: A web-based annotation tool for historical and other non-standard language data. In: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH). April 26, 2014, Gothenburg, Sweden, S. 86–90. URL: <http://aclweb.org/anthology/W/W14/W14-0612.pdf> [Stand: 23.03.2017].

- Bußmann, Hadumod (Hg.) (2008): Lexikon der Sprachwissenschaft. Mit 14 Tabellen. 4., durchges. u. bibliogr. erg. Aufl. Stuttgart.
- Dipper, Stefanie u.a. (2013): HiTS: ein Tagset für historische Sprachstufen des Deutschen. In: *Journal for Language Technology and Computational Linguistics* 28/1, S. 85–137. URL: http://www.jlcl.org/2013_Heft1/5Dipper.pdf [Stand 23.03.2017].
- Härd, John Evert (2000): Syntax des Mittelniederdeutschen. In: Werner Besch u. a. (Hgg.): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. 2. Teilbd. 2., vollst. neu bearb. u. erw. Aufl. (Handbücher zur Sprach- und Kommunikationswissenschaft 2.2). Berlin/New York, S. 1456–1463.
- Krause, Thomas und Amir Zeldes (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship in the Humanities* 31/1 (2016), S. 118–139. URL: <http://dsh.oxfordjournals.org/content/31/1/118> [Stand: 29.03.2017].
- Lasch, Agathe und Conrad Borchling (1956ff.): *Mittelniederdeutsches Handwörterbuch*. Fortgef. von Gerhard Cordes und Dieter Möhn. Neumünster/Kiel/Hamburg.
- Müller, Thomas u. a. (2015): Joint Lemmatization and Morphological Tagging with LEMMING. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 17–21 September 2015, Lisbon, Portugal, S. 2268–2274. URL: <https://aclweb.org/anthology/D/D15/D15-1272.pdf> [Stand: 23.03.2017].
- Pustejovsky, James und Amber Stubbs (2012): *Natural Language Annotation for Machine Learning. A Guide to Corpus-Building for Applications*. Beijing u. a.
- ReN-Team (2014) = Barteld, Fabian u. a.: *Transkriptionshandbuch – Auszug. Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)*. URL: <https://vs1.corpora.uni-hamburg.de/ren/media/pdf/transkriptionshb1.pdf> [Stand: 31.03.2017].
- ReN-Team (2016a) = Barteld, Fabian u. a.: *Annotationshandbuch. Teil 1: PoS und Morphologie. Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)*. URL: <http://www.referenzkorpus-mnd-nrh.de/media/pdf/posmorphhb1.pdf> [Stand: 23.03.2017].
- ReN-Team (2016b) = Barteld, Fabian u. a.: *Annotationshandbuch. Teil 2: Lemmatisierung. Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)*. URL: <http://www.referenzkorpus-mnd-nrh.de/media/pdf/lemmahb1.pdf> [Stand: 23.03.2017].
- Schiller, Anne u. a. (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Stuttgart/Tübingen. URL: www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [Stand: 29.03.2017].
- Schiller, Karl und August Lübken (1875–1880): *Mittelniederdeutsches Wörterbuch*. 6 Bde. Münster/Bremen.
- Schmid, Helmut und Florian Laws (2008): Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In: *Coling 2008. 22nd International Conference on Computational Linguistics*. Pro-

- ceedings of the Conference. Volume 1. 18–22 August 2008, Manchester, UK, S. 777–784. URL: http://dl.acm.org/ft_gateway.cfm?id=1599179&type=pdf [Stand: 31.03.2017].
- Schröder, Ingrid (2014): Das Referenzkorporus: Neue Perspektiven für die mittelniederdeutsche Grammatikographie. In: *Jahrbuch für germanistische Sprachgeschichte* 5, S. 150–164.
- Text Encoding Initiative (o. J.): URL: <http://www.tei-c.org/index.xml> [Stand: 29.03.2017].
- Zeldes, Amir (2016): ANNIS User Guide – Version 3.4.3. URL: http://corpus-tools.org/annis/resources/ANNIS_User_Guide_3.4.3.pdf [Stand: 05.04.2017].
- Zifonun, Gisela u. a. (1997): *Grammatik der deutschen Sprache*. Bd. 1 (Schriften des Instituts für deutsche Sprache 7.1). Berlin/New York.

