WiN
Wiedererzählen im Norden

Transkriptions- und Annotationshandbuch für das Pro*Niedersachsen-Projekt „WiN – Wiedererzählen im Norden. Digitale Analyse weltlicher Erzählungen in niederdeutschen Inkunabeldrucken“ Version 1.0

von

Chiara De Bastiani, Anabel Recker und Jan Christian Schaffert

herausgegeben von

Chiara De Bastiani, Marco Coniglio, Anabel Recker, Heike Sahm und Jan Christian Schaffert

Stand: September 2019

Gefördert aus Mitteln von

Niedersächsisches Ministerium für Wissenschaft und Kultur

GCDH

Dies ist das erste von mehreren Transkriptions- und Annotationshandbüchern, die im Verlauf des WiN-Projekts die Arbeitsweise dokumentieren und Änderungen festhalten werden. Die interdisziplinäre Zusammenarbeit zwischen Sprachwissenschaft, Literaturwissenschaft und Digital Humanities hat sich bisher als fruchtbar erwiesen, fordert die Forschung aber auch bei der Entscheidung für Methoden, Begrifflichkeiten und Darstellungskonventionen heraus. Wir möchten uns daher bei den Forscher*innen bedanken, die uns in diesem Prozess unterstützt und beraten haben.

Zunächst danken wir Michelle Disep, Niklas Foitzik, Karina Heepe, Sören Hoch, Susanne Müller und Carl Simon Spinger, den Studierenden des Seminars ‚Dracula und Brandan im Norden', deren Vorarbeiten uns den Einstieg in dieses Projekt erleichtert haben, sowie Svenja Walkenhorst, die als studentische Hilfskraft für uns tätig ist .

An der Entwicklung der technischen Abläufe und Planung der Annotationsroutinen war Thomas Krause von der HU Berlin maßgeblich beteiligt. Ihm und Martin Klotz danken wir für den freigebigen Einsatz ihrer Zeit und Expertise, um uns in computerlinguistischen Belangen zu beraten. Wichtige und hilfreiche Hinweise für die Planung der Annotation erhielten und erhalten wir weiterhin vom Hamburger Zweig des ‚Referenzkorpus Niederdeutsch/ Niederrheinisch' in Gestalt von Ingrid Schröder und Sarah Ihden. Auch Anne Breitbarth und Melissa Farasyn haben wir für linguistischen Input zu danken. Von literatur- und materialwissenschaftlicher Seite danken wir Falk Eisermann für zuverlässige Auskünfte in allen Inkunabelfragen.

Unser besonderer Dank gilt dem Niedersächsischen Ministerium für Wissenschaft und Kultur und dem Göttingen Center for Digital Humanities für die großzügige finanzielle Unterstützung.

Göttingen, September 2019

Chiara De Bastiani, Marco Coniglio, Anabel Recker, Heike Sahm und Jan Christian Schaffert

Inhalt

1. Das Korpus

Ziel des Projekts ist es, ein Vergleichskorpus frühneuhochdeutscher und mittelniederdeutscher Erzähltexte zu erstellen, das sechs bis acht kürzere gedruckte Erzähltexte weltlicher Thematik enthalten soll. Die Texte sollen ca. 3.000-5.000 Wörter lang und zwischen ca. 1480 und 1510 in beiden Sprachstufen gedruckt worden sein. Das Korpus verfolgt den Anspruch, die historisch-kritische Aufarbeitung der Frühdrucke[1] für die literaturwissenschaftliche Forschung mit den Anforderungen der historischen Sprachwissenschaft zu vereinen. Die Annotation soll ermöglichen, morphologische, lexikalische und syntaktische Erkenntnisse zu den Übersetzungsprozessen vom Hoch- ins Niederdeutsche zu gewinnen und von dieser Grundlage aus auf Arbeitsprozesse in den Offizinen des frühen Buchdrucks zu schließen.

In der Regel sind die frühneuhochdeutschen Fassungen zuerst entstanden und zu einem späteren Zeitpunkt ins Mittelniederdeutsche übertragen worden. Grundsätzlich handelt es sich bei den Texten um Erzählungen in Prosa, worin sich ein allgemeiner Trend der Literatur des späten Mittelalters wiederspiegelt. Die Prosa ist für die linguistischen Annotationen besser geeignet als gebundene Rede, da die Regeln der Syntax durch das Versmaß beeinflusst und Ergebnisse der syntaktischen Suchabfragen nicht verlässlich interpretiert werden können. Dennoch beinhaltet das WiN-Korpus zwei reimstrukturierte Texte: Die Schwankerzählung „Bruder Rausch“ ist auf Mittelniederdeutsch und auf Frühneuhochdeutsch als Reimpaargedicht erhalten. Der „Graf im Pflug“ basiert in der frühneuhochdeutschen Fassung auf einem Erzähllied des 15. Jahrhunderts, dessen Strophen im Druck markiert und durchgezählt sind. Bei der Übertragung ins Mittelniederdeutsche wird die Reimbindung aufgegeben. Diese beiden Beispieltexte ermöglichen einerseits den Vergleich von Vers- und Prosaübertragungen. Andererseits ist „Bruder Rausch“ neben den „Juden von Sternberg“ eine der wenigen weltlichen Kurzerzählungen, die vom Mittelniederdeutschen ins Frühneuhochdeutsche übersetzt wurden und somit als Grundlage einer Gegenprobe dienen können: Lassen sich für die Übertragung vom Frühneuhochdeutschen ins Mittelniederdeutsche die gleichen Strategien beobachten wie bei der Übertragung vom Mittelniederdeutschen ins Frühneuhochdeutsche? In der Korpusoberfläche können die fraglichen Texte später aus der Abfrage exkludiert werden, wenn Ergebnisse der syntaktischen Annotation nur für reinen Prosatext abgerufen werden sollen.

[1] Die Erscheinungsjahre der hier berücksichtigten Drucke (ca. 1480 bis ca. 1510) gehen über die im Gesamtkatalog der Wiegendrucke gesetzte Grenze von 1500 hinaus. Damit dies nicht zu Missverständnissen führt, verwenden wir statt der Begriffe ‚Wiegendruck‘ und ‚Inkunabel‘ den unschärferen Terminus ‚Frühdruck‘.

1.1 Die Texte

Dracula

Hdt.: N. N. Dracole Waida. Nürnberg 1488. Gedr. v. Marx Ayrer, GW 12524.

Ndt.: N. N. Dracole Waida. Lübeck 1488. Gedr. v. Bartholomäus Ghotan, GW 12531.

Die Juden von Sternberg

Hdt.: N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. v. Simon Koch, GW M44007.

Ndt.: N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. v. Simon Koch, GW M44009.

Der Graf im Pflug

Hdt.: N. N.: Der Graf im Pflug. Erfurt 1495. Gedr. v. Hans Sporer, GW 12575.

Ndt.: N. N.: Der Graf im Pflug. Magdeburg 1500. Gedr. v. Simon Koch, GW 12576.

Griseldis

Hdt.: Francesco Petrarca: Historia Griseldis. Augsburg 1471. Übers. v. Heinrich Steinhöwel, gedr. v. Güner Zainer, GW M31580.

Ndt.: Francesco Petrarca: Historia Griseldis. Hamburg 1502. Übers. v. Heinrich Steinhöwel, gedr. v. Drucker des Jegher, VD16 G3366.

Die Vier Kaufleute

Hdt.: N. N.: Die vier Kaufleute. Nürnberg um 1490. Gedr. v. Hans Hoffmann, GW 12616.

Ndt.: N. N.: Die vier Kaufleute. Lübeck 1495. Gedr. v. Matthäus Brandis, GW 12619.

Die sieben weisen Meister

Hdt.: N.N.: Historia septem sapientum Romae. Augsburg 1478. Gedr. v. Anton Sorg, GW 12858.

Ndt.: N.N.: Historia septem sapientum Romae. Lübeck 1478. Gedr. v. Lukas Brandis, GW 12873.

Bruder Rausch

Hdt.: N.N.: Bruder Rausch. Straßburg 1508. Gedr. v. Martin Flach d. J., VD 16 B8449.

Ndt.: N.N.: Broder Rusche. Stendhal 1488., Gedr. v. Joachim Westval, GW 12745.

1.2 Die Rechte

Die Nutzungsbedingungen der einzelnen Texte finden sich unter den jeweils angegebenen Links. Allgemein gilt: Sämtliche Texte sind in der Public Domain und können ggf. unter Angabe der den Text bereitstellenden Institution für nicht gewerbliche Zwecke verwendet werden.

Dracula Hdt. (Herzogin Anna Amalia Bibliothek):
https://haab-digital.klassik-stiftung.de/viewer/eula/

Dracula Ndt. (National Széchényi Library):
http://oszkdk.oszk.hu/help/hu/14b

Graf im Pflug Hdt. (Nationalbibliothek Berlin):
https://digital.staatsbibliothek-berlin.de/nutzungsbedingungen

Graf im Pflug Ndt. (Nationalbibliothek Berlin):
https://digital.staatsbibliothek-berlin.de/nutzungsbedingungen

Juden v. Sternberg Hdt. (Nationalbibliothek Berlin):
https://digital.staatsbibliothek-berlin.de/nutzungsbedingungen

Juden v. Sternberg Ndt. (Nationalbibliothek Berlin):
https://digital.staatsbibliothek-berlin.de/nutzungsbedingungen

Griseldis Hdt. (Nationalbibliothek Berlin):
https://digital.staatsbibliothek-berlin.de/nutzungsbedingungen

Griseldis Ndt. (Bayrische Staatsbibliothek):
https://www.bib-bvb.de/web/guest/datenschutzerklaerung-gateway-bayern
http://daten.digitale-sammlungen.de/~zend-bsb/impdat.php?w=1

Vier Kaufleute Hdt. (Nationalbibliothek Berlin):
https://digital.staatsbibliothek-berlin.de/nutzungsbedingungen

Vier Kaufleute Ndt. (Bayrische Staatsbibliothek):
https://www.bib-bvb.de/web/guest/datenschutzerklaerung-gateway-bayern
http://daten.digitale-sammlungen.de/~zend-bsb/impdat.php?w=1

Historia septem sapientum Romae Hdt. (Bayrische Staatsbibliothek):
https://www.bib-bvb.de/web/guest/datenschutzerklaerung-gateway-bayern
http://daten.digitale-sammlungen.de/~zend-bsb/impdat.php?w=1

Historia septem sapientum Romae Ndt. (Carl von Ossietzky Bibliothek Hamburg):
https://digitalisate.sub.uni-hamburg.de/nutzungsbedingungen.html
https://www.sub.uni-hamburg.de/rechtsvorschriften.html

Bruder Rausch Hdt. (Bayrische Staatsbibliothek):
https://www.bib-bvb.de/web/guest/datenschutzerklaerung-gateway-bayern
http://daten.digitale-sammlungen.de/~zend-bsb/impdat.php?w=1

Bruder Rausch Ndt. (Forschungsbibliothek Gotha):
https://archive.thulb.uni-jena.de/ufb/templates/master/template_ufb2/sites/terms.XML

2. Der Workflow

1. Auswahl des Textes

Die Auswahl der Texte folgt notwendigen und hinreichenden Bedingungen. Notwendige Bedingungen sind die Existenz und Verfügbarkeit des gleichen Textes auf Hoch- und Niederdeutsch. Der Bedingung des Drucks vor 1500 konnte in zwei Fällen nicht gerecht werden (hdt. Bruder Rausch und ndt. Griseldis). Hier stand die Verfügbarkeit eines vergleichbaren Textes über der arbiträren zeitlichen Definition der Frühdrucke.

Hinreichende Bedingungen sind zum einen die Gattungszuweisung Erzählung (die jedoch, bspw. durch die Juden von Sternberg, teilweise flexibel ist) und zum anderen die Länge, die im Schnitt 3.000-5.000 Wörtern nicht überschreiten sollte.

2. Beschaffung des Digitalisates

Im Regel- und Idealfall ist das Digitalisat online zugängig und in der Public Domain und Teil der Creative Commons. Existiert kein Digitalisat, werden Aufwand und Kosten einer Beschaffung geprüft und das Digitalisat ggf. in Auftrag gegeben.

3. Transkription in oXygen

Die Texte werden entsprechend den Transkriptionsrichtlinien der TEI sowie dem DTABf diplomatisch transkribiert. Die Wahl des Texteditors ist hierbei frei, jedoch hat sich der oXygen XML Editor als Standard des Projektes etabliert.

4. Normalisierung

Die diplomatische Transkription wird nach den unter 3.2 beschriebenen Regeln normalisiert. Die originale Interpunktion wird getilgt und durch eine moderne Interpunktion ersetzt.

5. Präeditierung

Aus den XML-Dateien werden alle Tags getilgt und die Interpunktionszeichen durch ein Leerzeichen vom vorangehenden Wort getrennt, um so eine störungsfreie Weiterverarbeitung in EXMARaLDA zu gewährleisten. Die XML-Dateien werden zu TXT-Dateien umgewandelt, die keinerlei Metadaten enthalten, damit sie in den Partitur Editor importiert werden können.

Ein weiterer wichtiger Schritt in der Präeditierungsphase ist die Kollationierung der zwei Fassungen des jeweiligen Textes. Die mittelniederdeutsche und die frühneuhochdeutsche Fassung werden verglichen und lexikalische, morphosyntaktische, syntaktische und textuelle Unterschiede gekennzeichnet. Dieser Schritt vereinfacht die Alignierung der Texte in dem Partitur Editor (vgl. 4.3).

6. Überführung nach EXMARaLDA

Jeder Text wird in der normalisierten präeditierten und der diplomatischen Version in EXMARaLDA überführt, und beide Versionen zu einer Datei kombiniert. Jede Annotation eines Textes besteht also zunächst aus zwei Dateien: der kombinierten frühneuhochdeutschen Version und kombinierten mittelniederdeutschen Version, jeweils bestehend aus diplomatischer und präeditiert-normalisierter Fassung. Nachdem diese TXT-Dateien für die Alignierung bearbeitet wurden, werden abschließend in einer einzelnen Datei vereint, sodass nun der Vergleich des frühneuhochdeutschen und des mittelniederdeutschen Textes in jeder Transkriptionsfassung möglich ist.

7. Alignierung

Im Partitur Editor werden die frühneuhochdeutsche und die mittelniederdeutsche präeditiert-normalisierte Transkription einander gegenübergestellt. Alle Wörter und zuvor durch Leerzeichen

abgetrennte Interpunktionen werden in Zellen realisiert, die fortlaufend nummeriert sind. Die Stellen, die in beiden Fassungen übereinstimmen, müssen sich nun in beiden Fassungen auf der Zellenebene gegenüberstehen (die Zellennummer spielt hierbei keine Rolle). Wenn eine Fassung zusätzlichen Text enthält, werden in der anderen Fassung so viele leere Zellen hinzugefügt, bis die Passagen an der Stelle, an der sie wieder übereinstimmen, sich auf der Zellenebene wieder gegenüberstehen. Diese Schritte werden auf den einzelnen Dateien getrennt durchgeführt, bis die Alignierung komplett ist. Danach werden die frühneuhochdeutsche und die mittelniederdeutsche Fassung der Texte in einer einzelnen Datei zusammengeführt. Diese Datei enthält nun vier Textebenen: Mnd [txt[2]], Mnd [dipl], Fnhd [txt], Fnhd [dipl]. Sie werden so angeordnet, dass die normalisierten Textversionen über eine align-Ebene aligniert werden können.

8. Kennzeichnung von übereinstimmenden Stellen

Auf der Alignierungsebene werden die übereinstimmenden Token nochmals durchnummeriert, da die vom Partitur Editor generierte Durchnummerierung der Zellen allein nicht ausreicht, um die maschinelle Verarbeitung in ANNIS zu ermöglichen. Damit die alignierten Stellen in ANNIS korrekt visualisiert werden können, müssen alle übereinstimmenden Token mit übereinstimmenden Zahlen gekennzeichnet werden. Abweichungen werden nicht durchnummeriert, sondern durch entsprechende Tags gekennzeichnet.

9. Auswertung der Alignierung

In der Ebene align_tag werden die Tags für die qualitative Auswertung der Alignierung hinzugefügt, die sich auf die Kollationierung des Textes beziehen. Diese erfassen lexikalische, morphosyntaktische, syntaktische und textuelle Unterschiede.

10. Lemma-Ebene

Jedem Wort in den Texten wird über die unter Punkt 4.5 genannten Wörterbücher ein Lemma zugewiesen.

11. pos-Tagging

Jedem Wort wird ein Part-Of-Speech-Tag (POS) zugewiesen. Unser Tagset bezieht sich auf die Tagsets STTS (Schiller et al. 1999), DDDTS und HiTS (Dipper et al. 2013), die für die Annotation von sprachlichen Korpora, die letzten zwei insbesondere für die historischen Korpora des Deutschen, entwickelt wurden. Ferner werden unter den Ebenen pos_punct_dipl und pos_punct_norm Interpunktionszeichen annotiert. Die Tags für die letzteren Ebenen wurden für unser Korpus entwickelt.

12. Satz-Ebene

Jeder Satz wird als eine Spanne annotiert. Das Tag für eine Satzspanne ist SU (Sentence Unit). Die kleinste mögliche SU besteht aus einem flektierten Verb; es wird nicht zwischen Haupt- und Nebensätzen unterschieden. Koordinierte Sätze werden als SU_Coord markiert, verschachtelte Sätze werden anhand von Unterstrichen und Indexierung getaggt, die die Teile des unterbrochenen Satzes anzeigen.

[2] Die txt-Ebene enthält die normalisierte Fassung.

3. Die Texterfassung

3.1 Die diplomatische Transkription

Die diplomatische Transkription erfolgt im XML-Format im oXygen-XML-editor. Die diplomatische Transkription versucht, den Druck hinsichtlich seiner Sonderzeichen, Kürzungen, Interpunktionszeichen, seines Layouts und Bildprogramms und seiner Strukturierung so genau wie möglich abzubilden. Im Einzelnen bedeutet dies, dass Varianten auf Graphemebene ebenso abgebildet werden, wie Kürzungs- und Absatzzeichen, Leerräume und punktuell auftretendes fremdsprachliches Textmaterial. Grundlage hierfür ist die Unicode-Schriftart Junicode.

3.2 Sonderzeichen und alternative Schreibweisen

Die Frühdrucke enthalten sowohl Sonderzeichen, die weder Buchstabe noch Ziffer sind, als auch alternative Grapheme und Superskripte, die sämtlich abgebildet werden müssen. Hierfür bietet sich der Schriftsatz Junicode an, mit dem diese Zeichen realisiert werden können.

1. Superskripte

Im Korpus treten nur Vokale als Superskripte auf. Sie werden ausschließlich mit Vokalen kombiniert, um deren Lautwert zu verändern:

Umlautung (im Mittelniederdeutschen ggf. Dehnung):	Diphthonge:
aᵉ, oᵉ, uᵉ	uᵒ

2. Alternative Grapheme

Als alternative Grapheme werden auf Graphemebene alternative Schreibweisen bezeichnet, die für den selben Lautwert verwendet werden können:

geschwänztes z	Schaft-s	Bogen-r	i und j ohne Punkt	y mit Trema
ʒ	ſ	ꝛ	ı ȷ	ÿ

3. Sonderzeichen: Kürzungszeichen

Kürzungszeichen weisen z. B. Auslassungen von Nasalen oder den Endungen –er oder –us aus:

Nasalstrich oder -tilde	er-Kürzel	us-Kürzel
¯ ~	‘	‘

Anmerkung: Für das –er und –us-Kürzel gibt es spezifische Formen, deren Codierung in Junicode möglich ist, jedoch immer wieder Visualisierungsprobleme verursacht. Daher werden beide Abkürzungen als ein einfaches Anführungszeichen transkribiert.

4. Sonderzeichen: Interpunktionen

Die Interpunktion der Texte besteht fast ausschließlich aus Punkten und Virgeln. Im Folgenden werden nur jene Interpunktionszeichen aufgelistet, die von modernen Interpunktionszeichen (wie Punkt oder Komma) abweichen.

Alinea-Zeichen	Mittelpunkt	Doppeltrennstich	Virgel
¶	·	⸗	/

Das Alinea-Zeichen markiert Absätze, der Mittelpunkt entspricht zumeist einem Punkt. Der Doppeltrennstrich wird analog zum Bindestrich verwendet und die Virgel bezeichnet eine Zäsur, die dem Punkt und dem Komma ähnlich ist.

3.3 Die Erfassung mit XML

Das Layout des Textes und seine Strukturierung in Seiten, Kapitel, Absätze und Überschriften wird als die Strukturierung der Information im Werk ebenso wie die Illustrationen gemäß der Richtlinien der Text Encoding Initiative und dem Basisformat des Deutschen Textarchivs in XML erfasst. Zusätzlich werden Metadaten erfasst, die Aufschluss über den Text selber geben.
Im Folgenden wird die Umsetzung in XML anhand der hierarchischen Struktur der XML dargestellt. Der folgende Abschnitt gibt demnach nicht nur Aufschluss über die verwendeten XML-Elemente und damit Markups, sondern verdeutlicht zugleich auch deren Struktur.

3.3.1 Die Metadaten

Die Metadaten der Quellen werden neben weiteren Angaben zu den Mitarbeiter*innen im sogenannten

<teiHeader>[3]
erfasst, speziell innerhalb des
<editionStmt>
<extent>
umfasst Informationen zum Umfang des Textes auf Image-, Token-, Typen- und Charakter-Ebene.
<sourceDesc>
umfasst in <bibl> zunächst allgemeingültige bibliographische Angaben der Form „Titel, Druckort, Erscheinungsjahr/Zeitraum, GW-Nummer, Bibliothek, Signatur“.
Diese Informationen werden im
<titleStmt> erfasst:
<author>
Der Autor des Werkes, wenn er sich ermitteln lässt (z. B. Francesco Petrarca für die Griseldis. Ansonsten N. N.).
<editor>
Der Drucker des Werkes. Wenn möglich wird der Drucker durch die Attribuierung des <editor>-Tags über den Katalog der deutschen Nationalbibliothek und die dort vergebene ID referenziert.
<publisher>

[3] Die XML-Commands sind Hyperlinks, die auf die jeweils zugrundeliegende Dokumentation des DTABfs oder, wenn diese nicht gegeben ist, der TEI verweisen.

Die Offizin, in der das Werk erschienen ist.

<msDesc> umfasst:

<repository>

Bibliothek oder Archiv, die oder das den Druck beherbergt.

<idno>

Signatur und URN des Digitalisates

<profileDesc> enthält zudem Informationen über die

<language>,

die Sprache des Textes, und der

<claszCode>,

der Informationen zur Textgattung enthält.

3.3.2 Die Textcodierung

Der Text wird wie folgt erfasst:

<Text> kann, je nach Vorgabe der Quelle, in drei Teile unterteilt sein:

<front> Titelblatt

<body> Fließtext

<back> Kolophon, wobei das Kolophon auch in den Fließtext integriert sein kann.

<front> kann folgende Elemente enthalten:

<pb>

Seitenumbrüche werden generell zu Beginn einer neuen Seite genannt
und verweisen sowohl auf die korrespondierende Scan-Seite, als auch auf die Foliierung der Quelle.

<titlepage>

Definiert die Titelseite und zeichnet mit dem Attribut type Haupttitel ("main") die Haupttitelseite aus. Andere Auszeichnungen, die vom DTABf erfasst werden können, werden nicht benötigt.

<titel part> Hier wird der Titel transkribiert.

<figure> Ermöglicht das Erfassen von Holzschnitten.

<note> Beschreibt die Holzschnitte, falls vorhanden.

<body> kann folgende Elemente enthalten:

<pb> (s. o.)

Erfasst geschachtelte Textabschnitte und gliedert diese durch das Attribut n="x", wobei x für die jeweilige Schachtelungsebene steht. n="1" umfasst beispielsweise den gesamten in <body> erfassten Fließtext, n="2" erfasst Kapitel innerhalb des Fließtextes.

<p>

Erfasst Absätze innerhalb der von erfassten Textabschnitte.

<lb/>

Erfasst Zeilenumbrüche am Ende jeder Zeile.

<opener>

Erfasst ggf. Einleitende Textpassagen im Fließtext, die daher nicht in <front>erfasst werden können.

<head>

Zeichnet Kapitelüberschriften aus.

<hi>

Markiert Hervorhebungen, die durch das Attribut rend="" spezifiziert werden.

<rs>

Referencing Strings werden im Corpus verwendet, um Eigennamen auszuzeichnen. Hier musste sich gegen die Konvention des DTAs entschieden werden, da diese mit <placeName> und <PersName> zu rigide Kategorien vorgeben, in die beispielsweise Namen von göttlichen Entitäten nicht integriert werden können. Gleiches gilt für die Notwendigkeit, Namen eindeutig als reale oder fiktionale Bezeichnungen zu definieren, was nicht immer klar zu entscheiden ist.

<hi>

Markiert Hervorhebungen, die durch das Attribut rend="" spezifiziert werden, z. B. Initialen.

<space>

Weist eine vom Drucker absichtlich positionierte Lücke im Text einer Zeile aus.

<supplied>

Weist von den Editoren ergänzten Text aus, der in der Quelle verloren oder unleserlich ist.

<fw>

Weist Bogensignaturen oder Kustoden aus. Diese werden mit dem Attribut type= und dem Wert "sig" für Bogensignaturen (Kustoden kommen im Korpus nicht vor) und dem Attribut place= , das die Werte "bottom" und "top" haben kann, spezifiziert.

<choice>

Weist Fehler aus, indem zunächst via:

<sic>

die in der Quelle überlieferte Schreibweise diplomatisch abgebildet wird und diese anschließend durch das Tag:

<corr>

korrigiert wird.

<back> (Kolophon)

3.3.3 Korrekturen, Fußnoten und ergänztes Textmaterial

Zu korrigierende Textstellen werden durch das Tag-Set <choice><sic></sic><corr></corr></choice> bearbeitet. Innerhalb des <sic>-Tags wird die Schreibung des Leittextes wiedergegeben, innerhalb des <corr>-Tags die korrigierte Schreibweise. Grammatische Fehler werden mit dem Tag <note></note> ausgewiesen. Da es sich hierbei um eine Fußnote handelt, muss der <note>-Tag entsprechend spezifiziert werden: <note place="foot"></note>. Ggf. können noch vorherige und folgende Fußnoten ausgewiesen werden.
Sicher zu ergänzendes Material wird durch das Tag <supplied></supplied> angegeben, unleserliche Stellen mit dem Tag <unclear></unclear> ausgewiesen. Über das Attribut @reason kann, wenn gewünscht, der Grund des Fehlens oder der Unleserlichkeit spezifiziert werden. Zeilen werden nur abschließend mit <lb/> beendet, um die Wohlgeformtheit der XML bei über Zeilenumbrüche verlaufenden Namen zu gewährleisten.

3.3.4. Auszeichnung von Sonderzeichen, Superskripten und alternativen Schreibweisen

Sonderzeichen, Superskripte und alternative Grapheme können entweder mit dem entsprechenden Junicode-Font direkt geschrieben oder per Code definiert werden. Jeder Code muss der Form &#xXXXX;

entsprechen, wobei XXXX das Sonderzeichen spezifiziert. ſ definiert beispielsweise ein ſ, ꝛ ꝛ und — einen Gedankenstrich. Im Rahmen des Projektes hat sich die Verwendung von Junicode als *best practice* herauskristallisiert. Da in den Junicode-Zeichen die Kodierung hinterlegt ist, kann sie vom/von der jeweiligen User*in und / oder Programm automatisch umgewandelt und so kompatibel gehalten werden.

3.4 Die normalisierte Transkription

3.4.1 s-, r- und z-Graphe

Zwischen alternativen Schreibweisen auf Graphie-Ebene wird nicht unterschieden und stets die moderne Schreibweise gewählt. Schaft-s (ſ) und Rund-r (ꝛ) werden durch <s> und <r> ersetzt, das geschwänzte z (ʒ) wird zu <z> normalisiert.

Beispiel:

Diplomatisch	Normalisiert
Der haiden ritter. reit aus vnd fraget die kriſtē leůte welichs deꝛ weg. gen mentʒ were	Der haiden ritter reit aus und fraget die kristen leute, welichs der weg gen Mentz were

N. N.: Der Graf im Pflug. Erfurt 1495. Gedr. v. Hans Sporer, GW 12575, Bl, 2r.

3.4.2 i-, j- und y-Graphe

In vokalischer Stellung werden <i> und <y> zu <i> normalisiert, <i> und <y> werden in konsonantischer Stellung durch <j> wiedergegeben, also *ia*, *yaer* oder *iammer* als *ja*, *jaer* oder *jammer* realisiert. Einschränkung: Sollte <j> für <i> verwendet werden, kann diese Normierung nicht weiter aufrechterhalten werden.

Beispiel J/j zu I/i:

Diplomatisch	Normalisiert
Jtem ein pꝛiſter genant her Peter Then iſt [...] vber eyn kommen: wy hernach volget	Item ein prister genant her Peter Then ist [...] uber ein kommen: wi hernach volget

N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. V. Simon Koch, GW44007, Bl. 2r.

Beispiel I/i zu J/j:

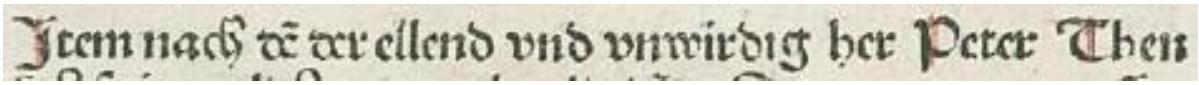
Diplomatisch	Normalisiert
Nach Criſti geburt. M. C C C C.lvi.iar	Nach Cristi geburt M CCCClvi jar

N. N. Dracole Waida. Nürnberg 1488. Gedr. v. Marx Ayrer, GW 12524, Bl. 1v.

3.4.3 u-, v- und w-Graphe

<u> für <v> und <v> für <u> werden entsprechend ihrem Lautwert ausgeglichen. Gleiches gilt für <w> für /u/ (z. B. in *iw* = *ju*; aber *iuw* = *juw*).

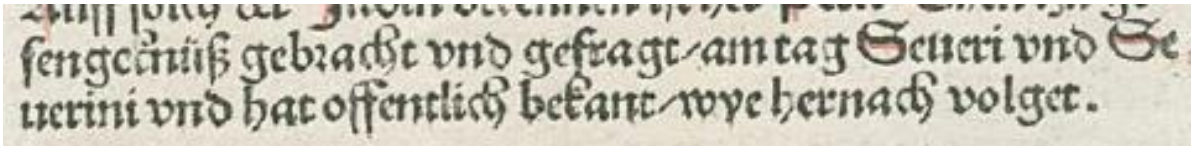
Beispiel V/v zu U/u:



Diplomatisch	Normalisiert
Jtem nach dē der ellend vnd vnwirdig her	Item nach dem der ellend und unwirdig her

N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. v. Simon Koch, GW44007, Bl. 2r.

Beispiel U/u zu V/v:



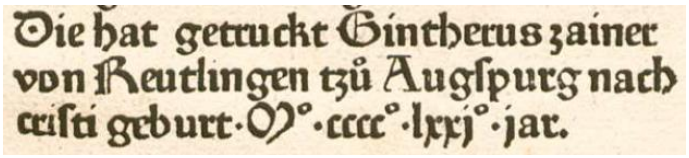
Diplomatisch	Normalisiert
[…] am tag Seueri vnd Se uerini vnd […]	[…] am Tag Severi und Se- verini und […]

N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. v. Simon Koch, GW44007, Bl. 2v.

Der niederdeutsche stimmhafte bilabiale Frikativ in intervokalischer Stellung [β] wie in niederdeutsch *ge**u**en, le**u**en, lo**u**en,* der auf graphematischer Ebene mit <u> oder <v> realisiert wird und dessen hochdeutsche Entsprechung ein Plosiv ist (z. B. <b>), wird zu <u> normalisiert. Im Auslaut, wie z. B. in *wiv*, bleibt das stimmlos realisierte <v> erhalten.

Für Zahlen gilt, dass Punkte vor und nach den (bisher ausschließlich römischen) Grundzahlen nicht wiedergegeben werden. Darüber hinaus werden Spatien zwischen den einzelnen Ziffern getilgt, um die Zahl als suchbare Einheit zu gestalten. Die oftmals als <j> oder <J> realisierte letzte Ziffer wird durch <i> oder <I> dargestellt. So wird z. B. *m. c c c c xcviiij.* zu *mcccxcviiii* normalisiert.

Beispiel Zahlen:



Diplomatisch	Normalisiert
Die hat getruckt Gintherus ȝainer	Die hat getruckt Gintherus Zainer
von Reutlingen tzů Augſpurg nach	von Reutlingen tzuo Augspurg nach
criſti geburt. M°.cccc°.lxxj°.jar	Cristi geburt Mcccclxxi jar

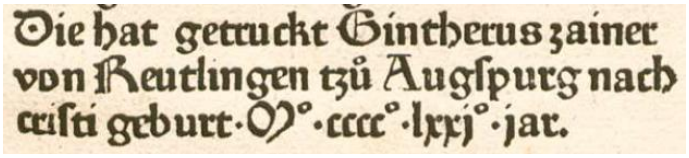
Francesco Petrarca: Historia Griseldis. Augsburg 1471. Übers. v. Heinrich Steinhöwel, gedr. v. Güner Zainer, GW M31580, Bl. 9r.

3.4.4 Abbreviaturen

Handelt es sich um Nasalstriche, *er-*, *-us* oder um *unde*-Kürzel, werden Abkürzungen stillschweigend aufgelöst. Der Tendenz im Frühneuhochdeutschen, durch Nasalstriche doppelte Konsonanten zu bilden, wird nicht nachgegangen. Ausgenommen sind Kürzungen von Doppelkonsonanten, wie z. B. *darumme* etc. Lateinische Abkürzungen werden stillschweigend entsprechend der gängigen Regeln aufgelöst. Vgl. hierzu ggf. Cappelli (1928).

3.4.5 Diakritische Zeichen

Überschriebenes <o> tritt nur über <u> auf und wird als <uo>, wie in *bruoder*, realisiert.
Beispiel:

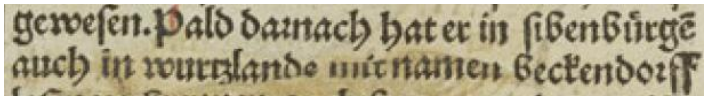


Diplomatisch	Normalisiert
Die hat getruckt Gintherus ȝainer	Die hat getruckt Gintherus Zainer
von Reutlingen tzuºAugſpurg nach	von Reutlingen tzuo Augspurg nach
criſti geburt. M°.cccc°.lxxj°.jar	Cristi geburt Mcccclxxi jar

Francesco Petrarca: Historia Griseldis. Augsburg 1471. Übers. v. Heinrich Steinhöwel, gedr. v. Güner Zainer, GW M31580, Bl. 9r.

Superskript-e über den Vokalen <a> und <u> wird hauptsächlich als Umlaut <ä> oder <ü> realisiert.

Beispiel:



Diplomatisch	Normalisiert
Pald daꝛnach hat er in ſibenburgē	Pald darnach hat er in Sibenbürgen
auch im wurtȝlande mit namen	auch im Wurtzlande mit namen

N. N. Dracole Waida. Nürnberg 1488. Gedr. v. Marx Ayrer, GW 12524, Bl. 1v.

Im Mittelniederdeutschen ist jedoch das parallel auftretende Dehnungs-e zu berücksichtigen, dass dem gedehnten Vokal nachgestellt wird, z. B. *Moesel* und nicht *Mösel*.

Beispiel:

Diplomatisch	Normalisiert
vn̄ darſulues alle latēn doeꝛſpeten	und darsulues alle laten doerspeten

N. N. Dracole Waida. Lübeck 1488. Gedr. v. Bartholomäus Ghotan, GW 12531, Bl. 1v.

Daher werden in den mittelniederdeutschen Texten alle superskribierten <e> über <o> als <oe> wiedergegeben.

3.4.6 Groß- und Kleinschreibung

Satzanfänge, *nomina sacra* und Eigennamen werden groß-, alle anderen Wörter – entsprechend der Überlieferung – kleingeschrieben. Dies bedeutete, dass ein unmotiviert großgeschriebener Wortanfang zugunsten der mehrheitlich verwendeten Form geändert wird.

3.4.7 Zusammen- und Getrenntschreibung

Bei Zusammen- und Getrenntschreibung folgt der Text der Vorlage, weicht jedoch dort ab, wo offensichtlich gegen die Silbengrenzen getrennt und so das Verständnis erschwert wird. Bei abgetrennten Präfixen muss im Einzelnen geprüft werden, ob eine getrennte Schreibweise im Frühneuhochdeutschen oder Mittelniederdeutschen belegt und der Trennstrich entsprechend obsolet ist. Fehlen am Ende einer Zeile die Trennungsstriche, so werden die Wörter nach der Gewohnheit des Drucks getrennt- oder zusammengeschrieben. Gibt es beide Möglichkeiten, so wird stets für die Getrenntschreibung entschieden. Unsinnige Zusammenschreibungen werden stillschweigend korrigiert.

3.4.8 Interpunktion

In der normalisierten Textfassung wird die originale Interpunktion getilgt und durch eine moderne Interpunktion ersetzt. Auf diese Weise bleiben beide Interpunktionen erhalten: die originale in der diplomatischen, die moderne in der normalisierten Textfassung.
Die Interpungierung der normalisierten Textfassung erfolgt behutsam. Hierbei soll es sich in erster Linie um eine die Struktur des Textes besser hervorhebende Maßnahme handeln, die letztlich auch eine Lese- und Verständnishilfe bietet und sich auf die notwendigsten Zeichen beschränkt, d. h. vor allem <.> und <,>. Zudem wird wörtliche Rede ausgezeichnet. Was eine Zeichensetzung indiziert, ist in manchen Fällen nicht eindeutig. Mit *unde* verknüpfte Passagen etwa müssen nicht zwangsläufig durch Interpunktion abgetrennt sein. Verlaufen sie aber über ganze Seiten, drohen sie, die Grenze zwischen Satz und Text zu verwischen, und erschweren das Leseverständnis. Schlussendlich entscheidet die/ der Bearbeiter*in nach eigenem Ermessen.

3.4.9 Zum Umgang mit Fehlern im Druck

In der diplomatischen Version werden die Fehler des Druckers abgebildet, in der normalisierten Version sind sie korrigiert. Wo der Setzer seinen Text missverstanden oder bereits mit einer missverstandenen

Vorlage gearbeitet hat, wird dies angemerkt. Grammatische Fehler werden nicht korrigiert, aber als solche ausgewiesen.
Fehlendes oder unleserliches Material wird, sofern eindeutig rekonstruierbar, ergänzt und als solches ausgewiesen. Kann das fehlende oder unleserliche Material nicht sicher bestimmt werden, wird dies angegeben und die Fehlstelle beibehalten.
Die XML-choice-Kommentare sollten in Zweitversionen der diplomatischen Transkriptionen bestehen bleiben (wie auch das Name-Tagging), da sie für die linguistische Annotationsarbeit momentan nicht gebraucht werden, aber später wichtig sind.

3.4.10 Die Präeditierung

Die XML-Dateien mit der normalisierten Fassung und diplomatischen Transkription der Drucke werden für die Überführung nach EXMARaLDA vorbereitet; die XML-Struktur wird zugrunde gelegt und alle Tags werden getilgt. Schließlich, werden die XML-Dateien zu TXT-Dateien umgewandelt, die nach EXMARaLDA überführt werden können. Interpunktionszeichen sowohl in der diplomatischen, als auch in der normalisierten Fassung werden durch ein Leerzeichen vom vorangehenden Wort getrennt. Das ermöglicht die Trennung von Wörtern und Interpunktionszeichen im Partitur Editor, und deren Erfassung als Token.
Der letzte Schritt in der Präeditierungsphase besteht aus der Kollationierung der mittelniederdeutschen und der frühneuhochdeutschen Fassung von jedem Text. Beide Fassungen werden in einer doc-Datei auf zwei Spalten verglichen, und Abweichungen werden in einer dritten Spalte annotiert. Die Fassungen werden verglichen, und lexikalische, morphosyntaktische, syntaktische und textuelle Unterschiede werden markiert. Ferner wird darauf geachtet, dass zwischen den zwei Fassungen in Anzahl von Wörtern und Struktur übereinstimmenden Passagen auf den zwei Spalten aligniert werden, sodass sich der/ die Annotator*in für die Alignierung im Partitur Editor danach richten kann.
Unter „übereinstimmenden Passagen“ werden die entsprechenden Textportionen in beiden Fassungen verstanden, die hinsichtlich der Anzahl von Wörtern, syntaktischer Struktur und allgemeiner Bedeutung gleich sind.

4. Die Annotation als Parallelkorpus

4.1 Die Überführung nach EXMARaLDA

Wenn die Präeditierung beider Versionen eines Textes erfolgt ist, wird die präeditierte normalisierte Version in das Format EXMARaLDA konvertiert. EXMARaLDA stellt das Tool Partitur Editor 1.6 zur Verfügung, mit dem die Annotation erfolgt. Die diplomatische Version wird in den Partitur Editor in eine separate Datei importiert, und mit der normalisierten Version zusammengeführt. Der Ausgangstext für die Alignierung ist die jeweilige normalisierte Fassung von dem frühneuhochdeutschen und dem mittelniederdeutschen Text, deswegen wird die diplomatische Fassung unter die diplomatische Fassung eingeführt.

Die mittelniederdeutsche und die frühneuhochdeutsche Version von einem Text bestehen nun aus zwei Partitur-Editor-Dateien, mit der jeweiligen normalisierten und diplomatischen Fassung. Diese zwei Dateien werden getrennt für die Alignierung bearbeitet und erst zusammengeführt, nachdem die Alignierung erfolgt ist.

4.2 Die Annotationsebenen

Über die Annotation werden sowohl metatextuelle Informationen als auch morphosyntaktische und syntaktische Merkmale kodiert. Außer den vier Textebenen (Mnd [txt],[4] Mnd [dipl], Fnhd [txt], Fnhd [dipl]) erhält das Parallelkorpus folgende Annotationsebenen:

- align: Die Alignierung setzt einzelne Wörter und Textstellen des Ausgangstextes in Beziehung zu ihren Übertragungen (vgl. 4.3).
- align_tag: auf dieser Ebene werden Abweichungen ausgewertet.
- lemma: Über die Lemmaebene wird jedes Wort einem Lemma zugeordnet (vgl. 4.5).
- pos (Wortart): Bei der pos-Annotation wird jedem Wort ein morphosyntaktisches Tag zugewiesen. Das Tagset basiert auf den Tagsets STTS (Schiller et al. 1999), DDDTS und HiTS (Dipper et al. 2013), die u.a. auch für das Referenzkorpus Althochdeutsch und das Referenzkorpus Mittelhochdeutsch verwendet wurden. Ferner werden auch Tags aus dem STTS verwendet. Unter 4.6 wird eine umfassende Beschreibung gegeben.
- pos_punct_dipl: Die Interpunktionszeichen in der diplomatischen Fassung werden auf dieser Ebene annotiert.
- pos_punct_norm: Die normalisierten Interpunktionszeichen werden auf dieser Ebene annotiert.
- sentence_chunk (Satzeinheit): Die Satzeinheiten (sentence unit = SU) werden als Satzspannen annotiert, die vom Anfang bis zum Ende der Satzeinheit verlaufen.
- facs_page (Verweis auf die Seite des Digitalisates): Verweist auf die Seite des Digitalisates, die der transkribierten entspricht. Es gilt: Die erste Seite des Digitalisates entspricht der ersten bedruckten Seite.
- page (editorische Paginierung): Sie beginnt mit der ersten bedruckten Seite des Frühdrucks l und wird fortlaufend gezählt.
- facs_paragraph (Abschnittsgliederung der Vorlage): Die originelle Abschnittsgliederung wird als Spanne annotiert.
- line (Zeile): Zeilenumbrüche werden als Spanne annotiert.
- comments (Anmerkungen zu einzelnen Textstellen): Die Kommentarebene enthält Kommentare der Annotator*innen, die für die Benutzer*innen des Korpus als nützlich erachtet werden. Im Fall unklarer Lesarten dokumentieren sie den Entscheidungsprozess oder verweisen auf Eigenheiten des Textes, die durch die Annotation nicht erfasst werden können.

4.3 Die Alignierung

In diesem Abschnitt wird das Alignierungsverfahren anhand der von Thomas Krause gegebenen Hinweise und der ersten Probeannotationen beschrieben; das Verfahren wird noch entwickelt und getestet[5]. Die Wörter, Sätze und Abschnitte des mittelniederdeutschen und des frühneuhochdeutschen Textes werden in zwei einzelnen Dateien verglichen und aligniert. Die diplomatische und die normalisierte Version der Texte kann jeweils gemeinsam angepasst werden (beide in zwei Ebenen in einer Datei). Man vergleicht die diplomatische und die normalisierte Ebene des frühneuhochdeutschen Textes mit der diplomatischen und normalisierten Ebene des mittelniederdeutschen Textes.
Der Partitur Editor zählt die Zellen fortlaufend. Ziel der Alignierung ist es zunächst, die Zellen so anzuordnen, dass Extratext (ET) in der einen Datei durch leere Zellen in der anderen Datei dargestellt wird. Durch das Hinzufügen von Zellen in der einen bzw. der anderen Datei erzielt man eine Entsprechung in

[4] Schreibweise, die in ihrer Form von den verwendeten Programmen diktiert wird.
[5] Insbesondere wird von Thomas Krause und Martin Klotz an der Humboldt Universität zu Berlin ein Modul von dem Transformationsprogramm SaltNPepper für die parallele Alignierung aufbereitet.

der Nummerierung der Zellen. Wenn ein Abschnitt eines Textes keine Entsprechung im Übertragungstext findet, sollte die Zellenzählung wieder übereinstimmen, sobald der Abschnitt endet. Erst diese händische Alignierung ermöglicht es, die vier Ebenen korrekt in einer Datei miteinander zu verbinden.
Die für die Alignierung verwendeten Ebenen werden im Folgenden abgebildet; die Beispiele sind aus unserer Annotation mit EXMARaLDA Partitur Editor 1.6 von dem Text „Die Juden von Sternberg“, in der mittelniederdeutschen und frühneuhochdeutschen Fassung (Gedr. v. Simon Koch, GW M44007, und gedr. v. Simon Koch, GW M44009) entnommen.
Auf Abbildung 1 wird die automatische von EXMARaLDA generierte Durchnummerierung der Zellen angezeigt; anhand von dieser automatischen Nummerierung werden der Ausgangstext und deren Übersetzung aligniert, sodass sich übereinstimmende Token derselben Zelle angeordnet werden.

	14 [00:	15 [00:14.0*]	16 [0	17 [0	18 [00:17	19 [0	20 [0	21 [00:2	22 [00:21.0*]	23 [00:	24 [00:23.0*
Mnd [txt]	Alle	man	si	to	wetten	,	de	grote	miszhande-linge	unde	oeveldath
Mnd [align]	1	2	3	4	5	6	7	8	9	10	11
Fnhd [align]	1	2	3	4	5	6	7	8	9	10	11
Mnd [align_tag]									LEX		LEX
Fnhd [align_tag]									LEX		LEX
Fnhd [txt]	Aller	menigklich	sei	tzu	wissen	,	der	grosz	misz-brauch	und	that

Abbildung 1 Die Nummerierung der Zellen

Die Alignierung erfolgt auf der normalisierten und diplomatischen Fassung von jedem Text. Die vier Textebenen werden so angeordnet, dass die normalisierten Textversionen in der Mitte über eine weitere Ebene aligniert werden können. Auf der Fnhd. [align] und der Mnd. [align] erfolgt die manuelle Alignierung der Token, indem die Ebenen die Token des Ausgangstextes nummerieren und mit einer parallelen Nummerierung auf deren Entsprechung in der Bearbeitung verweisen. Anstatt einen Text als Ausgangstext und den anderen als Übertragung zu annotieren, haben wir uns dafür entschieden, die Differenzen im Textbestand jeweils im Verhältnis zum anderen Text aufzunehmen. Auf diese Weise sollen die Unterschiede zwischen beiden Textfassungen so neutral wie möglich und jeweils durchsuchbar erfasst werden, ohne dass schon die Annotation durch die einer bestimmten Übersetzungstheorie geschuldeten Implikationen wertend beeinflusst ist. Die numerische Alignierung garantiert überdies die maschinelle Verarbeitung der Vergleichsergebnisse.
Der Ausgangstext und deren Übersetzung werden als [txt]-Ebene gekennzeichnet; ferner enthält jede Ebene Informationen über die jeweilige Sprache. Für die Bearbeitung werden der mittelniederdeutsche und der frühneuhochdeutsche Text so angeordnet, dass sich zwischen ihnen die align- und die align_tag-Ebenen befinden, wie man unter Abbildung 2 beobachten kann. Nach der Bearbeitung der Alignierungsebenen wird der übersetzte Text unter den Ausgangstext geschoben; dieser Schritt ermöglicht eine spätere benutzerfreundliche Visualisierung in ANNIS.

	14 [00:	15 [00:14.0*]	16 [C	17 [C	18 [00:17	19 [C	20 [C	21 [00::	22 [00:21.0*]	23 [00:	24 [00:23.0*	25 [C
Mnd [txt]	Alle	man	si	to	wetten	,	de	grote	miszhande-linge	unde	oeveldath	,
Mnd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Fnhd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Mnd [align_tag]											LEX	
Fnhd [align_tag]											LEX	
Fnhd [txt]	Aller	menigklich	sei	tzu	wissen	,	der	grosz	misz-brauch	und	that	,

Abbildung 2 Die [txt] Ebenen

Auf der mnd. [align] und der fnhd. [align] Ebenen werden übereinstimmende Token durchnummeriert; die Nummerierung erfolgt händisch. Diese Ebene wird nur annotiert, wenn komplette Übereinstimmung der Token vorhanden ist.

	14 [00:	15 [00:14.0*]	16 [C	17 [C	18 [00:17	19 [C	20 [C	21 [00::	22 [00:21.0*]	23 [00:	24 [00:23.0*	25 [C
Mnd [txt]	Alle	man	si	to	wetten	,	de	grote	miszhande-linge	unde	oeveldath	,
Mnd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Fnhd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Mnd [align_tag]											LEX	
Fnhd [align_tag]											LEX	
Fnhd [txt]	Aller	menigklich	sei	tzu	wissen	,	der	grosz	misz-brauch	und	that	,

Abbildung 3 Die [align] Ebenen

Auf der [align_tag] Ebenen erfolgt die qualitative Auswertung der Unterschiede zwischen dem Ausgangstext und deren Übersetzung. Diese Ebene wird nur annotiert, wenn Abweichungen vorhanden sind. Ein Beispiel für die manuelle Nummerierung der Token:

	14 [00:	15 [00:14.0*]	16 [C	17 [C	18 [00:17	19 [C	20 [C	21 [00::	22 [00:21.0*]	23 [00:	24 [00:23.0*	25 [C
Mnd [txt]	Alle	man	si	to	wetten	,	de	grote	miszhande-linge	unde	oeveldath	,
Mnd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Fnhd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Mnd [align_tag]											LEX	
Fnhd [align_tag]											LEX	
Fnhd [txt]	Aller	menigklich	sei	tzu	wissen	,	der	grosz	misz-brauch	und	that	,

Abbildung 4 Die [align_tag] Ebenen

Token in dem Ausgangstext werden progressiv auf der Alignierungsebene durchnummeriert; die parallele Durchnummerierung erfolgt auf der Alignierungsebene des übersetzten Textes, wie Abbildung 5 zeigt. Die Stellen, die keine Übereinstimmung aufweisen, werden nicht durchnummeriert, und die Nummerierung wird erst wieder aufgenommen, wenn Token übereinstimmen.

	14 [00:	15 [00:14.0*]	16 [C	17 [C	18 [00:17	19 [C	20 [C	21 [00::	22 [00:21.0*]	23 [00:	24 [00:23.0*	25 [C
Mnd [txt]	Alle	man	si	to	wetten	,	de	grote	miszhande-linge	unde	oeveldath	,
Mnd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Fnhd [align]	1	2	3	4	5	6	7	8	9	10	11	12
Mnd [align_tag]											LEX	
Fnhd [align_tag]											LEX	
Fnhd [txt]	Aller	menigklich	sei	tzu	wissen	,	der	grosz	misz-brauch	und	that	,

Abbildung 5 Die händische Alignierung

Abweichungen werden in der Präeditierungsphase mit der Kollationierung der Texte gekennzeichnet; es wird zwischen lexikalischen, morphosyntaktischen, syntaktischen und textuellen Abweichungen unterschieden. Die entsprechenden Tags werden auf der [align_tag] Ebene eingeführt. Textuelle Unterschiede, wie NC (No Correspondence) und ET (Extra-Text), korrelieren nicht mit der parallelen Nummerierung auf der [align] Ebene.
Im Folgenden wird das Tagset dargestellt, ferner sind alle Tags unter 7.3 aufgelistet.
Wenn ein Textabschnitt keine Entsprechung in einer Fassung des Textes aufweist, wird er mit NC gekennzeichnet. Das Tag wird als Spanne dargestellt, siehe Abbildung 6:

	0 [00:	1 [00:01.0*]	2 [00:02.0*]	3 [00:03	4 [00:04.0	5 [00:05.0*]	6 [0(	7 [00:07.0*	8 [00:08.	9 [00:09.0*]	10 [00	11 [00	12 [00	13 [(
Mnd [txt]	Van	der	misehandelinge	des	hi-ligen	sacramentes	der	boeszen	joeden	to	dem	Sternberge		.
Mnd [align]														
Fnhd [align]														
Mnd [align_tag]	NC													
Fnhd [align_tag]	NC													
Fnhd [txt]	Die	geſchicht	der	Jüden	tzum	Sternberg	ym	landt	tzu	Mecklenburg	.			

Abbildung 6 Der NC Tag

In Abbildung 6 werden beide [align_tag] Ebene mit dem NC-Tag gekennzeichnet; es handelt sich hier um einen speziellen Fall: Der gekennzeichnete Textabschnitt ist der Titel des Ausgangstextes und der Übersetzung. Wobei es sich hier um dieselbe Sinneinheit handelt, gibt es keine Berührungspunkte zwischen den zwei Fassungen; deswegen werden beide Ebenen mit NC getaggt.
Das Tag ET signalisiert, dass es in der jeweiligen Fassung, wo das Tag auftaucht, zusätzlicher Text als in der anderen Fassung gibt. Normalerweise wird dieses Tag mit NC in der anderen align_tag-Ebene kombiniert:

	33 [00	34 [00:33	35 [00:34.0*]
Mnd [txt]			
Mnd [align]			
Fnhd [align]			
Mnd [align_tag]	NC		
Fnhd [align_tag]	ET		
Fnhd [txt]	dem	tzarten	fronleichnam

Abbildung 7 Der ET Tag

Die Abweichungen auf der Wortebene oder in der Formulierung desselben Begriffs werden als LEX getaggt. Diese qualitative Auswertung der Abweichungen kann mit der Nummerierung der Token korrelieren, falls sie auf der Wortebene begrenzt sind:

	):	24 [00:23.0*
Mnd [txt]	:	oeveldath
Mnd [align]		11
Fnhd [align]		11
Mnd [align_tag]		LEX
Fnhd [align_tag]		LEX
Fnhd [txt]		that

Abbildung 8 Der LEX Tag

Morphosyntaktische Unterschiede werden mit dem Tag MOS gekennzeichnet:

	47 [00	48 [00
Mnd [txt]	to	dem
Mnd [align]	30	
Fnhd [align]	30	
Mnd [align_tag]	MOS	
Fnhd [align_tag]	MOS	
Fnhd [txt]	tzum	

Abbildung 9 Der MOS Tag

Abweichungen auf der Satzebene werden mit SYN getaggt. Wenn sich das SYN-Tag auf eine andere Satzart bezieht, werden beide [align_tag] Ebene als Spanne annotiert, während die [align] Ebenen leer bleiben:

	281 [0	282 [04:41.	283 [04:	284 [04:	285 [04:44	286 [04:	287 [04:46.0*]	288 [04:47.0*	289 [04:48.0*	290 [04:49.0*]
Mnd [txt]	unde	na	alle	dusser	smaheit	unde	miszhande-linge	des	hilgen	sacramentes
Mnd [align]										
Fnhd [align]										
Mnd [align_tag]	SYN									
Fnhd [align_tag]	SYN									
Fnhd [txt]	nach	gethoner	smach	von	den	Jüden	dem	heiligisten	sacrament	beschehen

Abbildung 10 Der SYN Tag

Wenn mehrere Konstituenten umgetauscht werden, werden sie als Spannen auf der [align] Ebene annotiert, die entsprechend nummeriert werden, während der ganze Satz als Spanne annotiert und mit dem Tag SYN auf der [align_tag] Ebene gekennzeichnet wird.

	292 [(	293	294 [04	295 [0	296 [04:5:	297 [04:5(	298 [04:57.0	299 [04:	300 [04:5'	301 [05	302 [05:01.0*]	303 [(
Mnd [txt]	heft	her	Peter	dat	wedder	umme	entfangen	van	Eleazar	des	joeden	wiff
Mnd [align]	158	159	160	161	1			2				
Fnhd [align]	158	159	160	161	2					1		
Mnd [align_tag]					SYN							
Fnhd [align_tag]					SYN							
Fnhd [txt]	hat	her	Peter	dasz	von	Eleazar	desz	Jüden	weib	wider	entpfan-gen	

Abbildung 11 Umgetauschte Konstituenten

Eine Textstelle kann mehrere Unterschiede aufweisen; in solchen Fällen, werden die verschiedenen Tags kombiniert:

	7 |	208 [03:	209 [	210 [0	211 |	212 |	213 [03:3	214 [03:33.0*]	215 [03:34.	216 |	217 [03:36.0*	218 [0
Mnd [txt]		Alszo	hir	na	in	der	joeden	bekantnisse	clerliken	uth	gesproken	wert
Mnd [align]					130	131	132	133				
Fnhd [align]					130	131	132	133				
Mnd [align_tag]		SYN_MOS_LEX_ETMND										
Fnhd [align_tag]		SYN_MOS_LEX_NCFNHD										
Fnhd [txt]		wi	dan	dasz	in	der	jüden	bekantnisz	lawt			

Abbildung 12 Kombinierung der Tags

In dieser Abbildung wird gezeigt, wie die verschiedenen Tags kombiniert werden. In der mittelniederdeutschen Fassung, ist mehr Text vorhanden; das wird als ETMND getaggt, während in der Fnhd. [align_tag] Ebene wird angezeigt, dass es keine Korrespondenz gibt. Das Tag ETMND erfolgt aus der Kombination von dem Tag „ET“ und der Bezeichnung der Sprache, „MND“.
Sollte ein*e Benutzer*in nur nach einer Art der Abweichungen suchen wollen, wird es durch die Verwendung von regulären Ausdrücken auf ANNIS möglich werden:

align_tag=/.*MOS.*/

Mit dieser Suchanfrage ist der/die Benutzer*in in der Lage, alle MOS Abweichungen zu finden, auch wenn sie mit anderen Tags kombiniert sind.[6]

Schließlich finden sich Stellen, an denen in einer Fassung des Textes ein Bild vorhanden ist, während in der anderen Fassung mehr Text zu beobachten ist, der dem Bild in der jeweils anderen Fassung entspricht. Solche Fälle werden in der [align_tag] Ebene mit den Tags ET und IMG angezeigt. Diese werden auf der jeweiligen Mnd [align_tag] und Fnhd [align_tag] Ebene kombiniert, sodass die Suche in beiden Richtungen ermöglichen wird, diese textuelle Eigenheit im Korpus zu finden.

4.4 Der Paratext und die Metadaten

In der Annotation mit Partitur Editor werden folgende Metadaten erfasst: Paginierung in der Vorlage, Abschnittsgliederung in der Vorlage, Zeilenumbrüche, editorische Paginierung.
Diese Metadaten werden als Spanne annotiert, die vom Anfang bis zum Ende der jeweiligen Einheit verläuft. Diese werden fortlaufend vom Anfang des Dokuments durchnummeriert. Zudem wird eine editorische Paginierung eingefügt, die vom ersten bedruckten Blatt und damit der ersten Seite des Digitalisates fortlaufend ist. Da auch die Foliierung entsprechend fortläuft, gilt: Bl. 1r = S. 1, Bl. 1v = S2, …, Bl Xr = S. X*2-1, Bl. Xv = S. X*2.

4.5 Die Lemmatisierung

Frühneuhochdeutsch:
Da der Lexembestand des Frühneuhochdeutschen lexikographisch bislang nicht vollständig erschlossen ist (angekündigte Fertigstellung des Frühneuhochdeutschen Wörterbuchs in 2027), wird der Wortbe-

[6] Vgl. Annis User Guide: Introduction. http://korpling.github.io/ANNIS/4.0/user-guide/index.html (zuletzt aufgerufen am: 22.06.2019). Es wird vorausgesetzt, dass man die Sprache abruft, in der die jeweilige Abweichung zu finden ist.

stand des Korpus anhand verschiedener Wörterbücher lemmatisiert, die hierarchisch geordnet konsultiert werden. Steht ein Lemma noch nicht im FWB_online, wird es nach dem FWB_gedruckt zugeordnet. Ist es dort ebenfalls nicht verzeichnet, erfolgt die Lemmatisierung nach dem DWB. Daraus ergibt sich folgende Hierarchie:

FWB_online > FWB > DWB

Mittelniederdeutsch:
Die Lemmatisierung des Mittelniederdeutschen berücksichtigt die aktuellen Ergebnisse, die an den großen Korpus- und Wörterbuchprojekten, dem Referenzkorpus Mittelniederdeutsch (ReN Team 2018) und dem ‚Mittelniederdeutschen Wörterbuch', erarbeitet werden. Die mittelniederdeutschen Texte des Korpus werden in Abstimmung mit den lexikographischen Standards dieser Projekte lemmatisiert, die Quellenhierarchie geht von Lasch/ Borchling/ Cordes/ Möhn (1956ff.) aus, ergänzt um Lübben/ Walther (1995):

LBCM > LW

Ein Ziel der Parallelannotation ist es, aufzuzeigen, wie genau Übersetzungsprozesse verlaufen und welche Mischungen zwischen mittelniederdeutschen und frühneuhochdeutschen Texten zu verzeichnen sind. Deshalb sind die Lemmata, denen ein Wort zugeordnet wird, nicht an die Ausgangssprache des Gesamttextes gebunden. Kommt in einem mittelniederdeutschen Text eine frühneuhochdeutsche Wortform vor, wird sie auch entsprechend dem frühneuhochdeutschen Lemma zugeordnet.

4.6 Die pos-Annotation

Das Tagset für das WiN-Projekt basiert auf den Tagsets DDDTS und HiTS, die u.a. auch für das Referenzkorpus Althochdeutsch und das Referenzkorpus Mittelhochdeutsch verwendet wurden. Ferner werden auch Tags aus dem STTS verwendet. Alle pos-Tags sind unter 7.1 und 7.2 aufgelistet. Den Wörtern wird keine weitere morphosyntaktische Annotation hinzugefügt, außer bei partizipialen Adjektiven. In diesem Fall wird wie folgt annotiert:

(1) der geschickt man

(N. N. Dracole Waida. Nürnberg 1488. Gedr. v. Marx Ayrer, GW 12524, Bl. 10r.)

Annotation: geschickt, ADJA < VVPP

Mit dieser Annotation wird signalisiert, dass das Adjektiv aus einem Partizip Präteritum abgeleitet ist.

Auf der normalisierten Textebene werden Worteinheiten aus der diplomatischen Version entweder zusammen- oder getrennt geschrieben. Dies wird gemäß der gängigen Praxis in den Wörterbüchern bestimmt. Falls zwei Wörter dann auf der Lemma-Ebene verbunden werden, und die Komponenten zwei verschiedenen grammatischen Kategorien angehören, stellt die pos-Annotation zwei getrennte Tags bereit. Dies ist zum Beispiel bei Verben mit trennbarer Partikel der Fall, oder bei komplexen Partikeln wie (2):

(2) *Uberein*kommen

(N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. v. Simon Koch, GW M44007, Bl. 2r)

Fnhd [txt]	uberein		kommen
Fnhd [dipl]	vber	eyn	kommen
Fnhd [lemma]	überein		kommen
Fnhd [pos]	PTKVZ		VVPP

Abbildung 13 Das Tagging einer komplexen Partikel

Wenn Artikel zu einer Präposition klitisiert werden, werden Artikel und Präposition auf der Lemma Ebene getrennt; auf der pos-Ebene, werden sie mit dem Tag APPRART getaggt:

(3) Ym lannd tȝu Mecklenburg

(N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. v. Simon Koch, GW M44007, 1r.)

Fnhd [txt]	g	in	dem	landt	tzu	Mecklenburg
Fnhd [dipl]	;	ym		landt	tȝu	Mecklenburg
Fnhd [lemma]		in	das	land	zu	Mecklenburg
Fnhd [pos]		APPRART		NA	APPR	NEO

Abbildung 14 Das Tagging von Präpositionen und klitisierten Artikeln

4.7 Die Annotation der Satzebene

Satzeinheiten werden als Spanne auf der sentence_chunk-Ebene annotiert;[7] sie werden als SU (Sentence Unit) gekennzeichnet. Die minimale Satzeinheit enthält ein finites Verb, wie im folgenden Beispiel zu beobachten ist:

(4) Aller menigklich [sey]$_{\text{FLEK.VERB}}$ tzu wissen/ der grosz myszbrauch und that.

(Ebd., 2r.)

[7] Die sentence_chunk-Tags sind unter 7.4 aufgelistet.

Alle Argumente des flektierten Verbes werden derselben SU zugeordnet, sowie Adjunkte, die das Verb oder seine Argumente modifizieren. Auch wenn Komplement- und Relativsätze Argumente von Verben oder Attribute zu Nomina realisieren, die sich im Hauptsatz befinden, werden sie als getrennte SUs annotiert (wie auch Adverbialsätze), da sie ein finites Verb aufweisen. In Abbildung 15 ist eine Satzeinheit annotiert:

	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Fnhd [txt]	Allermenigklich		sei	tzu	wissen		der	grosz	miszbrauch			und	that	
Fnhd [dipl]	ALler	menigklich	ſey	tʒuwiſſen		/	der	groſʒ	myſʒ=	bʁauch		vnd	that	.
Fnhd [lemma]	allermänniglich		sîn	zu	wissen		der	gros	Misbrauch			und	tât	
Fnhd [pos]	PI		VAFIN	PTKZU	VVINF		DDART	ADJA	NA			KON	NA	$.
Fnhd [pos_punct_dipl]						$/								$.
Fnhd [pos_punct_norm]														
Fnhd [sentence_chunk]	SU													
Fnhd [facs_paragraph]	2													
Fnhd [facs_page]	2r													
Fnhd [page]	3													
Fnhd [line]	3									4				
[comments]														

Abbildung 15 Satzeinheit

Es wird nicht zwischen Haupt- und Nebensätzen differenziert; in den älteren Sprachstufen ist die Wortstellung kein ausreichendes Kriterium zur Differenzierung zwischen Haupt- und Nebensätzen.

Koordinierte Sätze, die ein elidiertes Element enthalten können, werden als SU_Coord annotiert:

(5) Item ein prister genant her Peter Then ist mit den vorstockten Iüden uber eyn kommen: wy hernach volget **und hat yn tzwo hostien vorkaufft.**

(N. N. Die Geschichte der Juden von Sternberg. Magdeburg 1492. Gedr. v. Simon Koch, GW M44009, 2r.)

Das ausgelassene Subjekt in dem koordinierten Satz ist koreferent mit dem Subjekt in dem ersten Satz. Nicht nur Subjekte können von der Elision betroffen sein, sondern auch Objekte:

(6) Item na deme dat de elende unde unwerdige her Peter Dhen sick so iammerlick an dem hilgen sacrament vorgetten heft, den joeden **vorkoft** unde **overgeantwordet.**

(Ebd.)

Im Korpus finden sich auch komplex verschachtelte Sätze:

(7) [Item ein prister]$_a$ [**genant her Peter Then]** [ist mit den vorstockten Iüden uber eyn kommen]$_b$.

(Ebd.)

Wie man am Beispiel (7) sehen kann, ist der Hauptsatz (mit a und b gekennzeichnet) von einem verschachtelten Relativsatz unterbrochen. Solche Fälle werden anhand von Unterstrichen und Indexierung getaggt, die die Teile des unterbrochenen Satzes anzeigen.

	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124
Fnhd [txt]	Item	ein	prister		genant	her	Peter	Then	ist	mit	den	vorstockten		Jüden	uberein	kommen	
Fnhd [dipl]	Item	ein	prıſter		genant	her	Peter	Then	iſt	mit	den	voꝛ-	ſtockten	Iüden	vber	eyn	kommen
Fnhd [lemma]	item	ein	priester		nennen	herr	Peter	Then	sîn	mit	die	verstocken		Jude	überein	komnen	
Fnhd [pos]	FM	DIART	NA		VVPP	NA	NE		VAFIN	APPR	DDART	ADJA<VVPP		NA	PTKVZ	VVPP	
Fnhd [pos_punct_dipl]																	
Fnhd [pos_punct_norm]																	
Fnhd [sentence_chunk]	SU_1				SU_Intr				SU_1								
Fnhd [facs_paragraph]	4																
Fnhd [facs_page]																	
Fnhd [page]																	
Fnhd [line]	15												16				
[comments]																	

Abbildung 16 Das Tagging von einem unterbrochenen Satz

Die Tags SU_1 zeigen die zusammengehörigen Teile von dem Hauptsatz; das Tag SU_Intr signalisiert den verschachtelten Satz.
Titel, die kein Finitum aufweisen, werden auf der [sentence_chunk] Ebene als Title annotiert, und Titel, die ein Finitum aufweisen, als SU_Title.

5. Die Publikation in ANNiS

Nachdem der Annotationsprozess fertig ist, werden die Partitur-Editor-Dateien mithilfe von SaltNPepper 3.3.4 (Zipser und Romary 2010) in ANNIS3 konvertiert und für Suchabfragen zur Verfügung gestellt. Für die korrekte Konversion der Alignierung, wird von Thomas Krause und Martin Klotz ein spezifisches Modul und einen passenden Workflow in Pepper entwickelt.

Die Metadaten zu den Vorlagen werden in LAUDATIO (LAUDATIO II ,2019, DFG: https://gepris.dfg.de/gepris/projekt/189321318) abgespeichert.

6. Literaturangaben, Programme und Hilfsmittel

6.1 Wörterbücher und Hilfsmittel

Cappelli, Adriano: Lexicon Abbreviaturarum. Wörterbuch lateinischer und italienischer Abkürzungen wie sie in Urkunden und Handschriften besonders des Mittelalters gebräuchlich sind, dargestellt in über 14 000 Holzschnittzeichen. 2., verb. Aufl. Leipzig 1928.

DWB = Deutsches Wörterbuch von Jacob und Wilhelm Grimm. 16 Bde. in 32 Teilbänden. Leipzig 1854-1961. Quellenverzeichnis Leipzig 1971.

LBCM = Lasch, Agathe und Conrad Borchling: Mittelniederdeutsches Handwörter-buch. Fortgef. von Gerhard Cordes und Dieter Möhn. Neumünster/Kiel/Hamburg 1956ff.

LW = Mittelniederdeutsches Handwörterbuch v. August Lübben. Nach dem Tode des Verf. vollend. v. Christoph Walther. Darmstadt 1995.

P5 Guidelines for Electronic Text Encoding and Interchange Version 3.5.0. Last updated on 29th January 2019, revision 3c0c64ec4DTABf. URL: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (zuletzt aufgerufen am 20.06.2019).

DTA-Basisformat. Das von CLARIN-D und der DFG empfohlene TEI-Format für historische Text. URL: http://www.deutschestextarchiv.de/doku/basisformat/ (zuletzt aufgerufen am 20.06.2019).

DDDTS: Referenzkorpus Althochdeutsch. Dokumentation der Annotation: Tagsets clause. URL: https://www.deutschdiachrondigital.de/manual/tagsets/) (zuletzt aufgerufen am 22.06.2019).

HiTS: Dipper, Stefanie u. a.: HiTS: ein Tagset für historische Sprachstufen des Deutschen. In: Journal for Language Technology and Computational Linguistics 28/1 (2013), S. 85-137. URL: https://jlcl.org/content/2-allissues/9-Heft1-2013/5Dipper.pdf (zuletzt aufgerufen am 22.06.2019).

ReN: Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650). URL: https://vs1.corpora.uni-hamburg.de (zuletzt aufgerufen am 22.06.2019).

STTS: Schiller, Anne u. a.: Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart/Tübingen 1999. URL: https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html (zuletzt aufgerufen am 22.06.2019).

Schiller, Karl und August Lübben (1875–1880): Mittelniederdeutsches Wörterbuch. 6 Bde. Münster/Bremen.

6.2 Programme und Werkzeuge

SaltNPepper: Zipser, F. und Romary, L.: *A model oriented approach to the mapping of annotation formats using standards.* In: Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta http://hal.archives-ouvertes.fr/inria-00527799/en/ (zuletzt aufgerufen am 22.06.2019).

LAUDATIO: LAUDATIO II (2019) DFG https://gepris.dfg.de/gepris/projekt/189321318 (zuletzt aufgerufen am 22.06.2019).

ANNIS3: Krause, Thomas und Zeldes, Amir: *ANNIS3: A new architecture for generic corpus query and visualization.* In: Digital Scholarship in the Humanities 31 (2016). http://dsh.oxfordjournals.org/content/31/1/118 (zuletzt aufgerufen am 22.06.2019).

EXMARaLDA: Schmidt, Thomas und Wörner, Kai : EXMARaLDA: Creating, analysing and sharing spoken language corpora for pragmatic research. In: Pragmatics 19 (2009), S. 565-582. https://exmaralda.org/de/ (zuletzt aufgerufen am 22.06.2019).

7. Anhang: Tagsets

7.1. pos-Tagset (in Anlehnung an den HiTS, DDDTS und STTS Tagsets)

Attributives Adjektiv	ADJA	Determinativ, generalisierend, attributiv, vorangestellt	DGA
Prädikatives Adjektiv	ADJD	Determinativ, generalisierend, attributiv, nachgestellt	DGN
Attributives Adjektiv, nachgestellt	ADJN	Determinativ, generalisierend, substituierend	DGS
Substituierendes Adjektiv	ADJS	Determinativ, indefinit, attributiv, vorangestellt	DIA

Adjektiv, prädikativ oder adverbial, Teil eines Eigennamens	ADJDE	Determinativ, indefinit, artikelartig	DIART
Adjektiv, attributiv, Teil eines Eigennamens	ADJE	Determinativ, indefinit, prädikativ	DID
Adjektiv, ordinal, attributiv	ADJO	Determinativ, indefinit, attributiv, nachgestellt	DIN
Adjektiv, ordinal, attributiv, nachgestellt	ADJON	Determinativ, indefinit, substituierend	DIS
Adjektiv, ordinal, substantiviert	ADJOS	Determinativ, indefinit, (mit und ohne Determiner), negativ	DINEG
Adjektiv, substantiviert	ADJS	Determinativ, indefinit, negativ, nachgestellt	DINEGN
Präposition	APPR	Determinativ, indefinit, negativ, substituierend	DINEGS
Präposition mit Artikel	APPRART	Determinativ, possessiv, attributiv, vorangestellt	DPOSA
Postposition	APPO	Determinativ, possessiv, prädikativ	DPOSD
Adverb	AVD	Determinativ, possessiv, attributiv, nachgestellt	DPOSN
Adverb oder Konjunktion	AVD-KO*	Determinativ, possessiv, substituierend	DPOSS
Relativadverb, generalisierend	AVG	Sonderfall: Genitiv des PPER als DPOS	DPOS
Adverb, interrogativ	AVW	Determinativ, relativisch, substituierend	DRELS
Kardinalzahl, attributiv, vorangestellt	CARDA	Determinativ, interrogativ, attributiv, vorangestellt	DWA
Kardinalzahl, prädikativ	CARDD	Determinativ, interrogativ, prädikativ	DWD
Kardinalzahl, attributiv, nachgestellt	CARDN	Determinativ, interrogativ, substituierend	DWS
Kardinalzahl, substituierend	CARDS	Determinativ, interrogativ, relativ	DWREL
Determinativ, definit, attributiv, vorangestellt	DDA	Determinativ, interrogativ, substituierend, relativ	DWSREL
Determinativ, definit, artikelartig	DDART	Determinativ, (interrogativ,) generalisierend, attributiv	DWG
Determinativ, Definit/demonstrativ, prädikativ	DDD	Determinativ, (interrogativ,) generalisierend, relativ	DWGREL

Determinativ, definit/demonstrativ, attributiv, nachgestellt	DDN	Fremdsprachliches Material	FM
Determinativ, definit/demonstrativ, substituierend	DDS	Interjektion	ITJ
Determinativ, definit/ demonstrativ/relativ	DDREL	Konjunktion, neben- oder unterordnend	KO*
Determinativ, definit/demonstrativ, substituierend, relativ	DDSREL	Vergleichspartikel	KOKOM
Konjunktion, nebenordnend	KON	substituierendes Indefinitpronomen	PIS
Konjunktion, unterordnend	KOUS	attribuierendes Indefinitpronomen ohne Determinierer	PIAT
unterordnende Konjunktion mit "zu" und Infinitiv	KOUI	attribuierendes Indefinitpronomen mit Determinierer	PIDAT
Nomen Appellativum	NA	Pronomen, indefinit, aus Substantiv	PI
Eigenname	NE	Pronomen, indefinit, negativ	PINEG
Eigenname, Ort	NEO	Pronomen, personal, irreflexiv	PPER
Pronominaladverb, präpositionaler Teil	PAVAP	Pronomen, personal, reflexiv	PRF
Pronominaladverb, pronominaler Teil	PAVD	substituierendes Possessivpronomen	PPOSS
Pronominaladverb, pronominaler Teil, generalisierend	PAVG	attribuierendes Possessivpronomen	PPOSAT
Pronominaladverb, pronominaler Teil, interrogativ	PAVW	substituierendes Relativpronomen	PRELS
Pronomen, generalisierend	PG	attribuierendes Relativpronomen	PRELAT
Pronomen, indefinit	PI	Pronomen, reflexiv	PRF
Pronomen, interrogativ	PW	Pronomen, (interrogativ,) adverbial, generalisierend, relativ	PWGAVREL
Substituierendes Interrogativpronomen	PWS	Pronomen, (interrogativ,) generalisierend, relativ	PWGREL
Attribuierendes Interrogativpronomen	PWAT	Pronomen, interrogativ, relativ	PWREL

Pronomen, interrogativ, adverbial	PWAV	Pronominaladverb	PAV
Pronomen, interrogativ, adverbial	PWAVREL	Partikel bei Adjektiv oder Adverb	PTKA
Pronomen, (interrogativ,) generalisierend, auch Genitivattribut	PWG	Antwortpartikel	PTKANT
Pronomen, (interrogativ,) adverbial, generalisierend	PWGAV	Negationspartikel	PTKNEG
Verbzusatz	PTKVZ	Vollverb, finit	VVFIN
„zu“ vor Infinitiv	PTKZU	Vollverb, imperativ	VVIMP
Relativpartikel	PTKREL	Vollverb, infinitiv	VVINF
Auxiliar, finit	VAFIN	Vollverb, Partizip Präteritum, im Verbalkomplex	VVPP
Auxiliar, imperativ	VAIMP	Vollverb, Partizip Präsens, im Verbalkomplex	VVPS
Auxiliar, Infinitiv	VAINF	Infinitiv, Vollverb, substantiviert oder flektiert	VVINFS
Infinitiv, Auxiliar, substantiviert oder flektiert	VAINFS	Partizip Präteritum, Vollverb, attribuierend	VVPPA
Auxiliar, Partizip Präteritum, im Verbalkomplex	VAPP	Partizip Präteritum, Vollverb, prädikativ oder adverbial	VVPPD
Auxiliar, Partizip Präsens, im Verbalkomplex	VAPS	Partizip Präteritum, Vollverb, attribuierend, nachgestellt	VVPPN
Modalverb, finit	VMFIN	Partizip Präteritum, Vollverb, substantiviert	VVPPS
Modalverb, imperativ	VMIMP	Partizip Präsens, Vollverb, attribuierend	VVPSA
Modalverb, infinitiv	VMINF	Partizip Präsens, Vollverb, attribuierend, nachgestellt	VVPSN
Infinitiv, modal, substantiviert oder flektiert	VMINFS	Partizip Präsens, substantiviert	VVPSS
Modalverb, Partizip Präteritum, im Verbalkomplex	VMPP	Modalverb, Partizip Präsens, im Verbalkomplex	VMPS

7.2 Interpunktionstagsets

Diplomatische Interpunktionszeichen

Doppeltrennstrich	$-
Virgel	$/
Alinea-Zeichen	$p
Mittelpunkt	$m.
Doppelpunkt	$:

Normalisierte Interpunktionszeichen

Punkt	$.
Komma	$,
Semikolon	$;
Doppelpunkt	$:
Fragezeichen	$?
Ausrufezeichen	$!
Anführungszeichen	$“, $“

7.3 Alignierungstagset

NC	No Correspondence (Keine Entsprechung)
ET	Extra-Text
LEX	Lexikalischer Unterschied
MOS	Morphosyntaktischer Unterschied
SYN	Syntaktischer Unterschied
IMG	Abbildung

7.4 Sentence_chunk-Tagset

SU	Sentence Unit (Satzeinheit)
SU_Coord	Koordinierter Satz
SU_n	Teil von einem unterbrochenen Satz
SU_Intr	Eingenesteter Satz
Title	Titel ohne Finitum
SU_Title	Titel mit Finitum