Datenbank

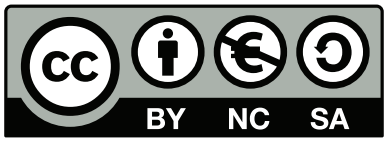
Deutsch V3 Diachron

Hauptsätze mit mehrfacher Vorfeldbesetzung in der Geschichte des Deutschen. Erarbeitet im Rahmen des Projekts *Lizenzierungsbedingungen für deutsche Verb-Dritt-Sätze in der Diachronie*, Projektnummer 76919537, Förderzeitraum 11/2017 bis 10/2020.

Annotationshandbuch

Carsten A. D. Dahlmann

`https://dv3d.uni-wuppertal.de`



Inhaltsverzeichnis

1 Einleitung

Das vorliegende Handbuch beschreibt den Aufbau der Datenbank DV3D (Deutsch V3 Diachron). Die Datenbank beinhaltet Sätze mit mehrfacher Vorfeldbesetzung, sogenannte Verb-Dritt-Sätze,[1] aus allen Überlieferungsstufen des Deutschen.
Die Daten wurden im Rahmen des DFG-Projekts *Lizenzierungsbedingungen für deutsche Verb-Dritt-Sätze in der Diachronie* (Projektnummer 76919537, Laufzeit 11/2017 bis 10/2020)[2] erhoben.

2 Vorbemerkungen

2.1 Struktur der Projektdatenbank

Diese Datenbank enthält Einzelsätze, die als Belege für *Hauptsätze mit Verbspäterstellung* in der Geschichte des Deutschen identifiziert wurden. Erfasst wurden Daten aus den Sprachstufen:

- Althochdeutsch
- Mittelhochdeutsch
- Frühneuhochdeutsch

Unter der Bezeichnung *Hauptsätze mit Verbspäterstellung* werden Hauptsätze verstanden, in denen dem Finitum mehr als eine Einzelkonstituente vorausgeht. Diese Abfolgen sind somit als Verletzungen der sog. Verbzweitregel im Hauptsatz aufzufassen, die für das Deutsche seit Beginn der Überlieferung relevant ist (vgl Axel 2007). Die bisherige Literatur

[1] Hier und im Folgenden ist mit V3 also stets eine Späterstellung (V3+) gemeint.
[2] Vgl. `http://gepris.dfg.de/gepris/projekt/376919537`

gibt Anlass zu Annahme, dass Hauptsätze mit Verbspäterstellung strukturell nicht einheitlich sind, sondern verschiedene syntaktische Verbstellungsmuster, also verschiedene Datensätze, verbinden. Dazu gehören einerseits Sätze, in denen das Finitum seine ursprüngliche Position in der rechten Satzklammer verlassen hat und nach links, in die linke Satzklammer bewegt wurde, wobei ein Vorfeld eröffnet wurde, das mehr als nur eine Einzelkonstituente beherbergt. Diese Sätze bezeichnen wir als *Verbdrittsätze*. Andererseits sind als Hauptsätze mit Verbspäterstellung Sätze zu bezeichnen, in denen die Voranstellung des Finitums unterbleibt, sog. *Verbletztsätze*. Die Diagnostiken zur Identifikation der Verbposition und zu der Klassifikation von Sätzen in die Kategorie der Verbdritt- bzw. Verbletztsätze werden in Kapitel 2.2.3 beschrieben.

Sätze, in denen diagnostische Elemente zur eindeutigen Identifikation der Verbstellung fehlen, sind strukturell ambig, können also sowohl als Verbdritt- als auch als Verbletztsätze interpretiert werden.

Die bisherige Forschung geht davon aus, dass sowohl Verbdritt- als auch Verbletztsätze über die gesamte Überlieferungsgeschichte des Deutschen bezeugt sind, Behagel (vgl. 1932) sowie Demske (2008) für eine Zusammenfassung der aktuellen Literatur.

Unsere Datenerhebungen haben die parallele Existenz diagnostischer Verbdritt- und Verbletztsätze nur im Althochdeutschen ergeben. Daher werden die Daten aus dem Althochdeutschen weiter nach Datensatz nebst Genre und lehnsyntaktischer Abhängigkeit differenziert. Dazu siehe Kapitel 2.2.4.

Die Daten aus dem Mittelhochdeutschen und Frühneuhochdeutschen werden nicht weiter nach strukturellen Datensätzen spezifiziert, sondern einheitlich in einem Subkorpus angegeben. Zum Umfang und zur Auswahl des gesichteten Datenmaterials, zur Datengewinnung und -klassifikation vgl. die jeweiligen Kapitel zum Mittelhochdeutschen (Kapitel 2.3) und Frühneuhochdeutschen (Kapitel 2.4).

2.2 Althochdeutsch

2.2.1 Analysierte Daten

Für die Analyse des Althochdeutschen wurde die komplette Überlieferung zugrundegelegt, die im digitalen *Referenzkorpus Altdeutsch* (ReA) (Donhauser et al.) zur Verfügung steht, wobei es sich entweder um übersetzte Prosatexte oder um Texte aus gebundener Sprache handelt (vgl. Fleischer 2006).
Die Daten wurden mittels geeigneter Suchabfragen an das ReA (in der Version 1.0) automatisch zusammengestellt und anschließend mit den weiter unten beschriebenen Diagnostiken manuell analysiert und kategorisiert.
Schließlich wurden die Daten mit den im Literaturverzeichnis (vgl. Kapitel 5.1: Editionen) angegebenen, einschlägigen Printeditionen abgeglichen. Nennenswerte Anpassungen mussten indes nur beim althochdeutschen *Tatian* vorgenommen werden, da dieser in ReA der Edition von Eduard Sievers, in unserem Falle jedoch der von Masser (1994) folgt, um die Beeinflussung des Zeilenwechsels auf die Verbstellung analysieren zu können.

2.2.2 Strukturelle Ambiguitäten

Für das Althochdeutsche kann nicht in allen Fällen davon ausgegangen werden, dass die für eine V2 respektive V3+ notwendige Bedingung der Verbbewegung im Hauptsatz stattgefunden hat. Um die Sätze zu klassifizieren, wurde sich entsprechender Diagnostiken (vgl. Kapitel 2.2.3) bedient, die somit zu einer Dreiteilung der althochdeutschen Daten führen:

- diagnostische V3-Sätze

- diagnostische Verb-Letzt-Sätze
- ambige Sätze

Ambige Sätze sind aufgrund von nicht ausreichendem Diagnosematerial nicht klar als V3 oder VL klassifizierbar. Sie sind somit als *ambig* kategorisiert und enthalten eine doppelt annotierte Ebene der topologischen Felder, jeweils einmal als potentieller V3-Satz, einmal als potentieller VL-Satz (vgl. Kapitel 3.3). Diese Datensätze spiegeln sich in der Datenbank in folgenden Subdatenbanken wider:

- *ahd_V3*
- *ahd_VL*
- *ahd_ambig*

Darüber hinaus werden die althochdeutschen Daten nach Differenz- und Korrespondenzbelegen differenziert, wenn es sich um Übersetzungstexte aus dem Lateinischen handelt, sowie nach Vers- und Prosatexten (vgl. Kapitel 2.2.4).

2.2.3 Diagnostiken

Entscheidend für die Kategorisierung eines Satzes als V3-Satz ist die Verbbewegung. Bei einem Satz wie (1) muss sichergestellt werden, dass das Verb in der Tiefenstruktur nicht an der Endposition steht und die ihm folgenden Konstituenten nicht durch eine Extraposition an ihre Stellung unterhalb des Verbs gelangt sind.

(1) her tho niouuiht antlingita imo
er da nichts antwortete ihm

‘Er antwortete ihm nichts.’
at ipse nihil illi respondebat. (Tatian 307, 25)

Dies ist über entsprechende Diagnostiken eruierbar: Diagnostisch für die Verbbewegung ist die Stellung des finiten Verbs links eines pronominalen Arguments (einschließlich eines Reflexivums) in der Linearisierung des Satzes, eines leichten Adverbs (ein- bis zweisilbig) oder einer Verbpartikel – also oberhalb von Konstituenten, die in OV-Sprachen nicht extraponierbar sind.
Auch für die Verb-Letzt-Stellung, die im Althochdeutschen noch in Hauptsätzen zu finden ist, existieren entsprechende Diagnostiken: Wenn das Finitum unmittelbar nach einem infiniten Verb, einer Verbpartikel oder einem prädikativen Nominal steht, handelt es sich um einen diagnostischen Verb-Letzt-Satz wie in (2).

(2) enti · ubil man · fona ubilemo horte · ubil fram bringit
und böser Mensch aus bösem Hort Böses hervor bringt

‘und der böse Mensch bringt von bösem Hort Übles hervor.’
et malus homo de malo thesauro profert mala
(Monsee Fragments 9, 19)

Indes gibt es im Althochdeutschen auch viele Sätze, die aufgrund des syntaktischen Materials nicht eindeutig als diagnostisch klassifiziert werden können – es handelt sich somit um ambige Sätze, wie z.B. (3).

(3) thara uuidar her thó quad
daraufhin wieder er da sprach

‘Daraufhin sprach er dann wieder’
at ipse ait (Tatian 119, 31)

Hauptsätze mit komplexen Verbformen, in denen das Finitum satzmedial erscheint, gelten als ambig, da die Abfolge Finitum – Infinitum nicht als diagnostisch für Verbbewegung gelten kann (z.B. bei Verb Raising und Verb Projektion Raising in heutigen OV-Varietäten, vgl. Bader/Schmid (2009)).

2.2.4 Zusätzliche Annotationsebenen

Um der Tatsache Rechnung zu tragen, dass jeder Beleg entweder aus einer übersetzten Prosaquelle oder aus einem Text mit gebundener Sprache stammt, wurden für das Althochdeutsche zusätzliche Annotationsebenen eingerichtet, die mithilfe einer AQL-Abfrage[3] aus der jeweiligen Subdatenbank herausgefiltert werden können:

(i) Differenzbelege
Daten aus den Übersetzungen sind mit dem dafür eingerichteten binären Tier *latSyn=yes/no* ausgestattet. Somit können sie von Belegen abgegrenzt werden, die potentiell von der lateinischen Syntax beeinflusst worden sind (Korrespondenzbelege). Beim althochdeutschen *Tatian* muss zudem bedacht werden, dass die Art und Weise der Übersetzung – jede Textzeile enthält nur so viel Text wie die lateinische Vorlage (vgl. Masser 1991: 92) – beeinflussen kann, wo das Verb steht. Aus diesem Grunde wurde für den *Tatian* die Ausgabe von Masser (1994) herangezogen. Zur Ebene *latSyn* vgl. auch Kapitel 3.6.2.

(ii) Gebundene Sprache
Um die Unterschiede der Lizenzierungsbedingungen von Versquellen und prosaischen Quellen aufzeigen zu können, können die jeweiligen Belege mit dem binären Tier *verse=yes/no* abgefragt werden.

[3] Für die Verwendung der Abfragesprache AQL vgl. `https://corpus-tools.org/annis/aql.html`.

(iii) Sonstige Ahd-Ebenen
Neben oben beschriebenen Ebenen kommen im Althochdeutschen noch die Ebene *latin* sowie bei ambigen Sätzen die Verdopplung der Felderanalyse mit hinzu. Vgl. hierzu Kapitel 3.3 sowie Kapitel 3.6.1.

2.2.5 Zusammenfassung Althochdeutsch

Somit ergibt sich für das Althochdeutsche theoretisch folgende Datenmatrix, die mithilfe der Subdatenbanken einerseits und der zusätzlichen Annotationsebenen andererseits abgefragt werden können – wobei faktisch für die Kombination *Korrespondenzbeleg* + *Prosa* keine Daten vorliegen.

Tabelle 1: Datenmatrix Althochdeutsch

V3, Diff, Prosa	VL, Diff, Prosa	Ambig, Diff, Prosa
V3, Diff, Vers	VL, Diff, Vers	Ambig, Diff, Vers
V3, Korr, Prosa	VL, Korr, Prosa	Ambig, Korr, Prosa
V3, Korr, Vers	VL, Korr, Vers	Ambig, Korr, Vers

Diese Datenstruktur existiert aufgrund der Besonderheiten der althochdeutschen Belege nur für diese Sprachstufe, nicht für das Mittel- oder Frühneuhochdeutsche.

2.3 Mittelhochdeutsch

Zum Mittelhochdeutschen wurden Prosatexte aus dem *Referenzkorpus Mittelhochdeutsch* (ReM) (Klein/Wegera et al. 2016) überprüft. Da über das ReM keine automatische Zusammenstellung mithilfe von geeigneten Suchabfragen möglich war, wurden die Daten manuell erhoben. Hierzu

wurden die folgenden diplomatischen Textausgaben ausgewählt und dabei auf eine ausbalancierte Verteilung über verschiedene Zeiträume und Dialektgebiete geachtet:

- *Bamb* – Bamberger Arzneibuch (um 1150, Rheinfränkisch)
- *Engelth* – Engelthaler Schwesternbuch (zwischen 1330 u. 1346, Ostfränkisch)
- *FR* – Franziskaner Regel (erste Hälfte des 14. Jh., Schwäbisch/ Alemannisch)
- *Klagschrift* – Klagschrift der Gesellschaft der alten Geschlechter zu Mainz (Rheinfränkisch-Hessisch, 1322)
- *Leipz* – Leipziger Predigten A (erste Hälfte des 14. Jh., Obermitteldeutsch)
- *Mill* – Millstätter Predigtsammlung (zweite Hälfte des 13. Jh., Bairisch)
- *MP* – Mitteldeutsche Predigten (1200, Rheinfränkisch)
- *Spec* – Speculum Ecclesiae (zweite Hälfte 12. Jh., Bairisch/ Alemannisch)
- *Wess* – Wessobrunner Glauben und Beichten (zweite Hälfte 12. Jh., Bairisch)

Da bei der Analyse der Daten keine diagnostischen Verb-Letzt-Hauptsätze gefunden werden konnten, ist davon auszugehen, dass stets eine Verbbewegung in die linke Satzklammer stattgefunden hat. Somit ist eine Aufteilung in Subdatenbanken wie im Althochdeutschen nicht notwendig. Die Belege werden somit einheitlich der Kategorie V3 zugeordnet (auch wenn es sich nicht ausschließlich um diagnostische V3-Fälle handelt).
Da die Belege ausschließlich entweder autochthonen Texten oder freien Übersetzungen aus dem Lateinsichen entnommen sind, ist auch eine Abfrage mit den Ebenen *latSyn* und *verse* hier nicht erforderlich.
Alle Belege finden sich somit in der Subdatenbank *mhd.*

2.4 Frühneuhochdeutsch

Die frühneuhochdeutschen Texte wurden ausschließlich aus dem digitalen *Bonner Frühneuhochdeutschkorpus* (Korpora.org) übernommen. Auch diese Daten wurden mangels möglicher Abfragemechanismen manuell analysiert.[4] Auch hierbei wurde auf eine ausbalancierte Verteilung über verschiedene Zeiträume und Dialektgebiete geachtet.
Es wurden die folgenden Texte des Korpus ausgewählt:

- *Dür. Chr.* – Johannes Rothe: Düringische Chronik des Johann Rothe, Text 253, thueringisch, 15. Jh.
- *H. Kottanerin* – Helene Kottanerin: Die Denkwürdigkeiten der Helene Kottanerin, Text 113, mittelbairisch, 15. Jh.
- *Main. Naturlehre* – Die sogenannte „Mainauer Naturlehre“ der Basler HS. B., Text 211, osthochalemannisch, 14. Jh.
- *Moscouia* – Sigmund Herberstein: Moscouia der Hauptstat der Reissen, Text 115, elsaessisch, 16. Jh.
- *Nuwe Boych* – Dat nuwe Boych. Zünfte und Bruderschaften, Text 151, ripuarisch, 14. Jh.
- *Pass. Mathesij* – Johannes Mathesius: Passionale Mathesij, Text 145, obersächsisch, 16. Jh.
- *Philander* – Hans Michael Moscherosch: Gesichte Philanders von Sittewald von Hans Michael Moscherosch, Text 237, 17. Jh.
- *Pillenreuth* – Eine neue Quelle für die Kenntnis des mystischen Lebens im Kloster Pillenreuth, Text 133, 15. Jh.

Auch in den erhobenen Daten des Frühneuhochdeutschen sind keine Verb-Letzt-Sätze gefunden worden.[5] Somit ist eine Unterteilung in Subdatenbanken auch hier nicht nötig.
Zum Frühneuhochdeutschen wurden ausschließlich Prosatexte unter-

[4] Mittlerweile ist das Bonner Frühneuhochdeutschkorpus auch über ANNIS abfragbar.
[5] Dass sie allerdings für das Frühneuhochdeutsche existieren, ist bezeugt. Vgl. Demske 2008.

sucht. Auch hierbei handelte es sich entweder um autochthone Texte oder freie Übersetzungen aus dem Lateinischen. Zusätzliche Annotationseneben sind folglich ebenfalls nicht notwendig.
Alle Belege finden sich somit in der Subdatenbank *fnhd*.

2.5 Technische Grundlagen

Die ermittelten Belege wurden mit *EXMARaLDA* annotiert und sind über eine Abfrage mit einer *ANNIS*-Instanz abrufbar (vgl. Kapitel 2.6.2)
EXMARaLDA ist ein ursprünglich für mündliche Korpora programmiertes System, das aus mehreren Komponenten, unter anderem einem Partitur-Editor, besteht und eine Datenannotation über mehrere Ebenen ermöglicht (vgl. Schmidt/Wörner 2014). Die in der DV3D-Datenbank verwendeten Ebenen werden in Kapitel 3 erläutert.
ANNIS ist ein Datenbanksystem zur Visualisierung solcher Mehrebenendaten (vgl. Krause/Zeldes 2016).
Ein Export von *EXMaRALDA*-Daten sowie anderen Ausgangsformaten und ein Import zu *ANNIS* sowie anderen Datenbankformaten kann mit den Corpus-Tools *Salt* und *Pepper* (vgl. Zipser/Romary 2010) erzielt werden.

2.6 Veröffentlichung

Veröffentlicht sind die reinen *EXMARaLDA*-Daten einerseits sowie die vorbereiteten *ANNIS*-Datenbanken andererseits. Veröffentlichungsort ist eine hierfür erstellte Zenodo-Community:
`https://zenodo.org/communities/dv3d/`

2.6.1 Die EXMARaLDA-Daten

Die *exb*-Daten enthalten noch Ebenen, die während des Projektzeitraums als Hilfsebenen zur Arbeit an den Daten genutzt wurden, jedoch nicht relevant für die Veröffentlichung sind. Hierbei handelt es sich um die Ebenen *clause* und *language* sowie um Ebenen mit Hilfsübersetzungen wie *transWord* und *transClause* – außerdem teils nie genutzte Ebenen wie *tos*, *foc* und *tos*.

2.6.2 ANNIS-Datenbanken

Die *EXMARaLDA*-Daten wurden schließlich mit den für die Veröffentlichung ausgewählten, relevanten Ebenen (vgl. Kapitel 3) in fünf Datenbanken im ANNIS-Format exportiert:

- *ahd_V3*
- *ahd_VL*
- *ahd_ambig*
- *mhd*
- *fnhd*

Diese fünf Datenbanken[6] können zur Abfrage mit AQL entweder in eine lokale ANNIS-Instanz importiert werden – hierzu müssen der *ANNIS-Kickstarter*,[7] eine *Java*-Laufzeitumgebung sowie eine lokale *PostgreSQL*-Datenbank installiert sein – oder, falls aktuell verfügbar, über eine zur Verfügung gestellte Online-Instanz bei Projekten wie LAUDATIO[8] aufgerufen werden.

[6] In dieser Dokumentation *Subdatenbanken* oder *Subkorpora* genannt.
[7] Der ANNIS-Kickstarter ist auf der Corpus-Tools-Homepage zu finden: `https://corpus-tools.org/annis/download.html`
[8] `https://www.laudatio-repository.org/`

Da das der DV3D-Datenbank zugrundeliegende Projekt selbst eine permanente Verfügbarkeit einer Online-Instanz nicht garantieren kann und auch keinen Einfluss darauf hat, welche Datensammlungen die ANNIS-Datenbanken in eine Online-Instanz importieren, werden, sofern bekannt, auf der DV3D-Webseite entsprechende Online-Instanzen verlinkt: `https://dv3d.uni-wuppertal.de`

2.6.3 Die Lizenz

Die veröffentlichten Daten sind unter einer *Creative Commons „Namensnennung – Nicht- kommerziell – Weitergabe unter gleichen Bedingungen 4.0 International"-Lizenz* (CC BY-NC-SA) veröffentlicht.[9]
Die Wahl dieser restriktiven Lizenz beruht auf der Abhängigkeit von den Lizenzen und Nutzungsbedingungen des Referenzkorpus Altdeutsch (CC BY-NC-SA) und des Bonner Frühneuhochdeutschkorpus, das eine Nutzung ihrer Textsammlung für wissenschaftliche Zwecke ohne kommerzielle Nutzung erlaubt.[10] Da nur das Referenzkorpus Mittelhochdeutsch unter der freien CC BY-Lizenz veröffentlicht ist, wurde auf eine Aufsplittung der veröffentlichten Daten verzichtet und CC BY-NC-SA als kleinster gemeinsamer Nenner gewählt.

2.7 Zitierweise

Wenn Sie das Korpus zitieren möchten, verwenden Sie bitte die folgende Zitation:

Svetlana Petrova, Nicholas Catasso, Carsten A. D. Dahlmann & Christopher Saure (2021). Deutsch V3 Diachron. Hauptsätze mit mehrfacher

[9] `https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode`
[10] Vgl. `https://korpora.zim.uni-duisburg-essen.de/FnhdC/`.

Vorfeldbesetzung in der Geschichte des Deutschen (Version 1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4743882

Wenn Sie im Speziellen dieses Handbuch zitieren wollen: Carsten A. D. Dahlmann (2021). Deutsch V3 Diachron. Hauptsätze mit mehrfacher Vorfeldbesetzung in der Geschichte des Deutschen. Annotationshandbuch. Zenodo. https://doi.org/10.5281/zenodo.4744338

3 Annotationsebenen

Das DV3D-Projekt enthält Annotationsebenen in den folgenden Kategorien:

- Belegebenen
  - *edition*
  - *text*
  - *lemma*
- Morphosyntaktische Ebenen
  - *pos*
  - *inflection*
  - *gf*
- Topologie (*field*)
- Informationsstruktur (*is*)
- Referenzebenen
  - *document*
  - *position*
  - *verse*
- Komparatistische Ebenen
  - *latin*

- *latSyn*

Einige der Ebenen wurden von den digitalen Korpora *Referenzkorpus Altdeutsch* (ReA) (Donhauser et al.), *Referenzkorpus Mittelhochdeutsch* (ReM) (Klein/Wegera et al. 2016) und vom *Bonner Frühneuhochdeutsch-Korpus* (Korpora. org) übernommen. Die übernommenen Elemente sind im jeweiligen Unterkapitel erläutert.

Die Annotationsregeln dieser Ebenen werden im Folgenden beschrieben.

3.1 Belegebenen

Aufgrund der Variation in der Schreibung sind zwei Ebenen notwendig, die den reinen Quelltext wiedergeben. Die einschlägigen Korpora differenzieren ebenfalls bereits alle zwischen einer diplomatischen Textwiedergabe und einer normalisierten. Dieses Schema wird hier übernommen. Die Ebenen sind in den verschiedenen Referenzkorpora indes unterschiedlich benannt. DV3D vereinheitlicht diese Ebenen für alle Sprachstufen in *edition* für die quellnahe Wiedergabe und *text* für die normalisierte Variante.
Darüber hinaus wird in der Ebene *lemma* eine lemmatisierte Wiedergabe der annotierten Wörter hinzugefügt, die auf den jeweiligen Lemma-Ebenen der Referenzkorpora beruht.

3.1.1 edition

Diese Ebene gibt den Text so nahe an der Quelle wie möglich wieder. Grundlage ist die jeweilige quellnahe Ebene der drei Referenzkorpora, wobei die Schreibweise beim Althochdeutschen mit den im Literatur-

verzeichnis (6.1: Editionen) angegebenen Printedition abgeglichen wurde.

- ahd: *edition* → *edition*
- mhd: *tok_dipl* → *edition*
- fnhd: *anno::gefunden* / *anno::gelesen* → *edition*

Sofern es möglich ist, das jeweilige Schriftzeichen mithilfe von Unicode darzustellen, geschieht dies. Somit enthält *edition* alle in den Editionen dargestellten diakritischen Zeichen, Ligaturen und im Neuhochdeutschen nicht mehr verwendete Grapheme (wie das lange s <ſ>) sowie besondere satzinterne Interpunktionszeichen wie Hochpunkte. Nasalstriche und Abkürzungszeichen werden ebenfalls quellnah übernommen und erst in der Ebene *text* aufgelöst und normalisiert.

(4) *edition:* ze aller heiligē me**ſſ**e an e**í**ner naht do […]
text: ze aller heilig**en** me**ss**e an e**i**ner naht do […]

Engelth, 14_1-ofrk-PU-G > M406-G1; 79a, 13-15

Getrennt- und Zusammenschreibung, die den tatsächlichen Wortgrenzen widerspricht, wird in *edition* dennoch so wie in der Quelledition dargestellt und erst in der Ebene *text* normalisiert.

(5) *edition:* Alſo hiute **andem** achtodim tâge . lîe er ſich beſnîden
text: also hiute **an dem** achtodim tage lie er sich besniden

Spec, 12_2-bairalem-PV-G > M214-G1; 12ra, 13-14

Auch die Groß- und Kleinschreibung wird, inklusive Binnenmajuskeln, von der Quelle übernommen.

(6) *edition:* [...] da gieng ich zu der edelen kung**I**nn
text: [...] da gieng ich zu der edelen kung**i**nn

H. Kottanerin 32, 18-19; 113

Wörter und Abkürzungen, denen ein anderes Alphabet zugrunde liegt (wie beispielsweise die an das griechische Alphabet angelehnte Schreibweise des Namens *Christus*), werden in lateinischer Schrift wiedergegeben, wenn sich die Quelledition ebenfalls lateinischer Grapheme bedient – also <x> und <p> anstatt <χ> und <ρ>.

(7) a. *edition:* [...] **xpān** [...] · uueset bitten te
text: [...] **kristan** [...] uueset bittente

Monsee Fragments 68, 6

b. *edition:* Dar nach do die **xPnheit** ge=wůchs [...]
text: darnach do die **kristenheit** gewuochs [...]

Leipz, 14_1-omd-PV-X > M536-No; 139rb, 9-11

Nicht abgebildet wird die satzabschließende Interpunktion, da es erstens häufig strittig ist, wo der Satz endet, zweitens eine Darstellung eines solchen Endzeichens eine Interpretation diesbezüglich seitens des Projekts bedeuten würde und drittens schließlich weil die Datensätze häufig sowieso nur einen Teil des Gesamtsatzes abbilden.
Im althochdeutschen *Tatian*, der in DV3D nach Masser (1994) zitiert wird, sind der Ebene *edition* die Zeilenbrüche, markiert mit <//>, hinzugefügt.
Die Datensätze enthalten stets den analysierten Matrixsatz und, sofern diesem ein Nebensatz folgt, die jeweilige satzeinleitende Konjunktion. Befindet sich ein eingeschobener Nebensatz vor dem Finitum, ist die-

ser in der Regel mit annotiert; nur bei sehr langen Unterbrechungen des Matrixsatzes werden Auslassungspunkte verwendet.

3.1.2 TEXT

Diese Ebene ist die normalisierte Fassung von *edition*. Sie ist nicht von den Referenzkorpora übernommen, sondern folgt den DV3D-eigenen Normalisierungsregeln. Diese Nomalisierungsregeln sind wie folgt:
Für eine bessere Auffindbarkeit von Textstellen über die Suchmaske ist vor dem Hintergrund der in den Quellen unterschiedlich dargestellten Groß- und Kleinschreibung hier ausschließlich in Kleinschreibung transkribiert worden. Das gilt auch für Eigennamen.
Die Ebene enthält ausschließlich Schriftzeichen, die in der normierten neuhochdeutschen Rechtschreibung kodifiziert sind. Alle darüber hinausgehenden diakritischen Zeichen und Ligaturen sowie Nasalstriche und Abkürzungszeichen werden aufgelöst und transkribiert. Hierbei werden auch die im Mittel- und Frühneuhochdeutschen hochgestellten <e> über Vokalen als eigenes Graphem transkribiert (also <ue> statt <ü>), um keine Interpretation hinsichtlich einer realisierten Monophthongisierung vorzunehmen.
Die Graphien von <u>, <v>, <w> respektive <uu> und <vv> werden nicht normalisiert, sondern ebenfalls dem Quelldokument entsprechend transkribiert.
Syntaktische Grapheme wie Siglen, Hochpunkte sowie weitere satzinterne Punkte etc. werden ausgelassen. Damit bleibt nach der oben ausformulierten Regel, nur Zeichen in *text* darzustellen, die in der neuhochdeutschen Rechtschreibung kodifiziert sind, nur noch das Komma als transkribiertes satzinternes Interpunktionsgraphem übrig.

(8) *edition:* **D**arumb we**n̄** man f**ů**r ain gefangnen bitt / spricht Er
text: **d**arumb we**nn** man f**ue**r ain gefangnen bitt spricht er

Moscouia, C4v., 1-2; 115

Getrennt- und Zusammenschreibung, die in *edition* die neuhochdeutschen Wortgrenzen bricht, ist hier normalisiert. Dies betrifft auch Verschmelzungsformen. Somit werden ab der Ebene *text* die normalisierten Wortgrenzen nach unten in die übrigen Ebenen weitervererbt.[11]

(9) a. *edition:* [...] ſo **wider** **sagich** dem tiuuile [...]
text: [...] so **widersag ich** dem tiuuile [...]

Wess, 12_2-bair-P-X > M098-N1; 2a, 17-19

b. *edition:* [...] vnd **vpme** huse saissen [...]
text: [...] vnd **vp me** huse saissen [...]

Nuwe Boych 440, 23-26; 151

Die in *edition* dargestellten Nasalstriche und andere Abkürzungsmarkierungen sind transkribiert. Dies gilt auch für aus dem Griechischen abgeleitete Grapheme:

(10) a. *edition:* [...] **xpān** [...] · uueset bitten te
text: [...] **kristan** [...] uueset bittente

Monsee Fragments 68, 6

b. *edition:* Dar nach do die **xƤnheit** ge=wůchs [...]
text: darnach do die **kristenheit** gewuochs [...]

Leipz, 14_1-omd-PV-X > M536-N0; 139rb, 9-11

[11] Technisch ist hierfür in den alt- und mittelhochdeutschen Subdatenbanken die Ebene *lemma* zuständig (vgl. Kapitel 3.1.3).

3.1.3 lemma

Die Ebene *lemma* entspricht den jeweiligen Lemma-Ebenen der Referenzkorpora.

- ahd: *Lemma* → *lemma*
- mhd: *Lemma* + *LemmaLemma* → *lemma* – wobei die Trennung von Adverbien und Partikelverben aus Gründen der Einheitlichkeit wieder aufgehoben wurde (Vgl. Klein/Dipper 2016: 12-15).
- fnhd: *lemma* → *lemma* – wobei nur die Lemmata übernommen wurden, die in der HTML-Ansicht des Korpus ein zugewiesenes Lemma aufweisen.

Die Ebene *lemma* ist zudem in den alt- und mittelhochdeutschen Subkorpora als zweiter *EXMARaLDA*-Sprecher festgelegt. Da von *lemma* aus wieder größere Ebenen mergen können, enthalten die anschließend nach *ANNIS* importierten Daten daher keinerlei Ebenen-Dopplungen in der Darstellung. Beim *fnhd*-Korpus wurde mangels einer vollständigen *lemma*-Ebene auf diesen zusätzlichen Sprecher verzichtet.

3.2 Morphosyntaktische Ebenen

3.2.1 pos (part of speech)

Die sehr fein gegliederten *pos*-Tags sind angelehnt an das *DDDTS*-Tagset des *Referenzkorpus Altdeutsch (DDD – Deutsch Diachron Digital)*. Dieses wiederum beruht auf dem Stuttgart-Tübingen-Tagset, dem STTS (Schiller et al. 1999), und wurde für die Verwendung bei alten Sprachstufen angepasst.
Die Referenzkorpora des Mittel- und Frühneuhochdeutschen verwenden

das HiTS (Historisches Tagset) (Dipper et al. 2013), das ebenfalls am STTS angelehnt ist, sich jedoch in nicht geringem Maße vom DDDTS unterscheidet. Um eine einheitliche Annotierung in DV3D zu gewährleisten, wurden die *pos*-Daten einheitlich auf Grundlage des DDDTS annotiert, das hier jedoch einige wenige Abweichungen und Ergänzungen erfahren hat (vgl. Kapitel 4.2.1).
Somit wurde auch die im Mittelhochdeutschen vorgenommene Trennung von Adverbien und Partikelverben (Vgl. Klein/Dipper 2016: 12-15) aus Gründen der Einheitlichkeit wieder aufgehoben.

(11) a. *tok_dipl:* dar nach
pos: PAVD PAVAP

diplomatische Quellannotation und *pos* im ReM

b. *edition:* dar nach
text: darnach
pos: ADV

Editionsebene, Normalisierung und *pos* in DV3D

Die Ebene *pos* gibt die morphologisch-funktionalen Eigenschaften des jeweiligen Wortes wieder. Kodiert sind hierbei, im Einklang mit der Verfahrensweise der Referenzkorpora, sowohl die Wortart (wie beispielsweise *V* für Verb) als auch die morphologische Form (z.B. *FIN* für *finit* oder *INF* für *infinit*) als auch, falls vorhanden, eine syntaktische Funktion, wenn beispielsweise ein Adjektiv substantiviert ist und somit substituierend (*S*) verwendet wird. Zudem wird auf dieser Ebene die satzinterne Interpunktion annotiert.

(12) *edition:* dea · ubilun · auuar · uurphun · uz
pos: DDA $(ADJS $(ADV $(VVFIN $(PTKVZ

Monsee Fragments 15, 20

Im Gegensatz zu den großen Referenzdatenbanken findet in der DV3D-Datenbank ausschließlich eine beleg-bezogene *pos*-Annotierung statt und keine zusätzliche Annotierung des lemma-bezogenen *pos*. Die im Tagset vorgesehenen Detailbezeichnungen gewährleisten die Unterscheidung von Ausgangs- und Zielwortart in einem Tag, beispielsweise durch Tags wie *VVPPA*, was ein adjektivisch verwendetes Partizip Perfekt beschreibt.

(13) *edition:* endi dhiu **chiborgonun** hort dhir ghibu
pos: KON DDA **VVPPA** NA PPER VVFIN

Isidor 7, 3

3.2.2 INFLECTION

Die Ebene *inflection* bildet die Flexionseigenschaften von Verben und Nomina ab. Sie ist nur für das Alt- und Mittelhochdeutsche annotiert, wo die grammatische Bestimmung vom jeweiligen Referenzkorpus übernommen wurde.

- ahd/mhd: *inflection* → *inflection* – wobei die Genera aus *inflectionClass* ergänzt sind und die Flexionsbezeichnungen dem hier punktuell abweichenden, eigenen DV3DTS-Schema entsprechen (vgl. Kapitel 4).

(14) *text:* unde aber doh petont sie in
inflection: IND_PRES_PL_3 MASC_PL_NOM_3 MASC_SG_ACC_3

N Psalmen 69, 21

3.2.3 GF (GRAMMATICAL FUNCTION)

Die Ebene *gf* stellt die Satzglieder der annotierten Phrasen dar. Die *gf*-Annotation findet stets aus Perspektive des Matrixsatzes statt. Nebensätze sind somit als Ganzes in ihrer Funktion im Matrixsatz annotiert. Diskontinuierliche Phrasen erhalten an beiden Stellen das gleiche Tag.
Die Annotation von Attributen wird vermieden; bei erweiterten Konstituenten wird keine genaue Auflösung vorgenommen: Wenn ein Attribut nicht innerhalb seiner übergeordneten Kategorie, sondern getrennt auftritt, wird beim getrennten Attribut die grammatische Funktion der übergeordneten Kategorie annotiert.

(15) *edition:* Aña aber des Küig Ludwigē schwester ist verheirat […]
gf: **SU** A **SU** FIN PART

[Anna] aber **[des Königs Ludwig Schwester]**, wird verheiratet […]

Moscouia, D4r., 33-34; 115

Analytische Verbformen erhalten verschiedene Tags für ihre jeweiligen Bestandteile.
Eine vollständige Liste der *gf*-Tags findet sich in Kapitel 4.2.3.

Hier und im Folgenden sollen nun für die DV3D-eigenen Ebenen jeweils kurze Abfragebeispiele die Einführung der Ebenen abrunden.
Im folgenden Abfragebeispiel wird ein Satz gesucht, in dem unmittelbar vor dem Finitum ein Adverbial und ein Subjekt stehen, die in der Reihenfolge Adverbial > Subjekt auftreten.

(16) gf="A" & gf="SU" & gf="FIN" & #1.#2 & #2.#3

3.3 Topologie

Die Ebene *field* unterteilt den Satz gemäß des topologischen Feldermodells. Bei den ambigen Sätzen des Althochdeutschen sind zwei Ebenen für die topologischen Felder vergeben: einmal als *field-V3* für die Annahme, der betreffende Satz sei V3, einmal als *field-VL* für die Annahme, er sei VL. Da jede der beiden Analysen die einschlägige sein könnte, werden in der Datenbank beide angeboten.

(17)

text:	tu sia	lazest	ersterbendo
field-V3:	Pre	Lsb	Middle
field-VL:	Middle	Rsb	Post

DeCons 2, 66, 25

Neben der linken und rechten Satzklammer und den anderen üblichen topologischen Feldern werden noch drei weitere Feldkategorien angenommen: IF (Illokutionsfeld), DL (Discourse Linker) und DF (Dislocation Field) (vgl. Kapitel 4.2.4).

Im folgenden Abfragebeispiel wird ein Satz gesucht, in dem das topologische Feld *Vorfeld* (*Pre*) ein Subjekt und ein Adverbial enthält und diese in der Reihenfolge *Subjekt* > *Adverbial* auftreten.

(18) field="Pre" & gf="SU" & gf="A" & #1_i_#2 & #1_i_#3 & #2.#3

3.4 Informationsstruktur

Die Informationsstruktur ist jener Bereich der sprachlichen Repräsentation, der die formalen Eigenschaften von Äußerungen in Abhängigkeit vom temporären Wissensstand der Diskursteilnehmer (vgl. Chafe 1976) bzw. vom gemeinsamen Wissenshintergrund (*common ground*) der Diskurspartizipanten zu einer bestimmten Phase im Diskurs (vgl. Krifka 2008) abbildet. Nach allgemeiner Auffassung ist die informationsstrukturelle Gliederung einer Aussage ein komplexes Phänomen, das gleichzeitig auftretende Differenzierungen auf den Ebenen des Informationsstatus (*Gegeben – Neu*), der prädikationellen Gliederung (*Topik – Kommentar*) und der *Fokus-Hintergrund*-Gliederung erfasst (vgl. Molnár 1993, Krifka 2008). In fortlaufenden monologischen Texten ist jedoch die lückenlose und eindeutige Zuweisung von Kategorien im Rahmen der *Topik-Kommentar-* sowie der *Fokus-Hintergrund*-Gliederung oft nicht möglich (vgl. Petrova/Solf 2009). Darüber hinaus ist die Identifikation dieser Kategorien bei der Annotation von natürlichen Texten (keinen diagnostisch gestalteten Frage-Antwort-Sequenzen) insgesamt sehr fehleranfällig. Aus diesem Grund wird eine andere Erfassung der betreffenden Phänomene präferiert, die im Rahmen der Taxonomie *gegebener* vs. *neuer* Information verbleibt. Annotiert wird somit nur der *Informationsstatus* von referierenden Ausdrücken, Ausdrücken also, die sich auf Individuen im fortlaufenden Diskurs beziehen und daher eine sichere Zuweisung der Kategorien *Gegeben* vs. *Neu* erlauben. Ein derartiges Modell wendet die Arbeit von Winkler (2017) zur Untersuchung der mehrfachen Vorfeldbesetzung im Gegenwartsdeutschen an.
Nicht Teil des Informationsstatus sind *Frames*, also temporale oder lokale Domänen, in der eine Proposition verankert wird, sowie *Sets* als Restkategorie.

In der Ebene *is* werden somit *Frames*, *Sets* und Satzthemen annotiert, wobei Letztere doppelt bestimmt werden, einmal hinsichtlich ihrer Vorerwähnung (*Thema* vs. *Rhema*) und einmal hinsichtlich ihrer *Salienz*, al-

so ob der Referent durch eine Vorerwähnung im Bewusstsein des Lesers präsent ist. Für die Analyse dessen wurde der jeweilige Kontext der zehn vorausgehenden Sätze manuell überprüft.
Adverbiale Bestimmungen hingegen wurden dahingehend überprüft, ob es sich um *Frames* (lokale oder temporale Domänen) oder *Sets* handelt.
Präpositionalobjekte und Argumentsätze wurden grundsätzlich als *Sets* deklariert.
Das Verb und alle seine analytischen Bestandteile wurden nicht annotiert.
Alle weiteren, oben nicht genannten grammatischen Funktionen wurden ebenfalls in die Restkategorie *Set* sortiert.
Die sich daraus ergebende Tag-Matrix ist in Kapitel 4.2.5 zu finden.

(19) *text:* her thô quad zi In
is: Th+sal Frame Set

Tatian 46, 27

Die jeweilige *is*-Einheit orientiert sich an den in *gf* analysierten Phrasen. Dies bedeutet, dass beispielsweise eine diskontinuierliche Phrase in *gf* ebenfalls als eine Einheit in *is* behandelt wird.

(20) *text:* uuala nu auh **huues** mac **dhesiu stimna** uuesan
gf: ILL A A **SU** FIN **SU** INF
is: Set Frame Set **Set** **Set**

‘**[Die Stimme wessen]**$_{gf:SU,\ is:Set}$ kann dies wohl nun auch sein’

Isidor 12, 5

Im folgenden Abfragebeispiel wird ein Satz gesucht, in dem die informationsstrukturelle Reihenfolge *Frame* > *salientes Rhema* > *salientes Thema* vorkommt.

(21) is="Frame" & is="Rh+sal" & is="Th+sal" & #1.#2 & #2.#3

3.5 Referenzebenen

3.5.1 document

Die Ebene *document* enthält eine Bezeichnung oder Sigle, die auf den Text verweist, in dem der Beleg enthalten ist. Die Siglen für das Althochdeutsche sind mit Abgleich des Literaturverzeichnisses (5.1: Editionen) selbsterklärend; die Bezeichnungen fürs Mittel- und Frühneuhochdeutsche können den Kapiteln 2.3 und 2.4 entnommen werden.

3.5.2 position

Die Ebene *position* verweist auf die konkrete Belegstelle.

Die althochdeutschen Positionsangaben beruhen auf den Printeditionen (vgl. Literaturverzeichnis (6.1: Editionen)).

In den mittelhochdeutschen Positionsangaben ist der Name des digitalen Pfades – bestehend aus der ANNIS-Gruppe und dem Dokument (vgl. auch Klein/Dipper 2016: 5) – sowie anschließend, nach dem Semikolon, Seite und Zeile der Textstelle im Originalmanuskript abgelegt.

(22) *document:* Wess
position: 12_2-bair-P-X > M098-N1; 2a,17-19

Text: *Wessobrunner Glaube* u. Beichte > ANNIS-Gruppe: *12_2-bair-P-X*, Dokument: *M098-N1*; Manuskript- bzw. Folioseite: *2a*, Zeile *17-19*

In den frühneuhochdeutschen Positionsangaben sind Seite und Zeile der Textstelle im Originalmanuskript abgelegt, zusätzlich, zur besseren Auffindbarkeit, ist nach dem Semikolon die Nummer des Textes angegeben.[12]

(23) *document:* Dür. Chr.
position: 12, 22; 253

Text: *Düringische Chronik*, Seite *12*, Zeile *22*; Textnummer *253*

3.5.3 verse

Die Ebene *verse* enthält den binären Wert *yes/no* und gibt an, ob es sich bei einem Text um einen in Verse gebundenen handelt. Dies betrifft nur das Althochdeutsche (vgl. Kapitel 2.2.4).

Die folgende Beispielabfrage sucht nach einem Satz in Versform, der ein Personalpronomen im Mittelfeld enthält:

(24) verse="yes" & field="Middle" & pos="PPER" & #1_i_#2 & #2_i_#3

3.6 Komparatistische Ebenen

Die folgenden Ebenen betreffen ausschließlich die althochdeutsche Subdatenbank.

[12] Die Textnummern sind die des *Bonner Frühneuhochdeutschkorpus*. Vgl. `https://korpora.zim.uni-duisburg-essen.de/annis/`

3.6.1 latin

Die Ebene *latin* bildet die lateinischen Originalvorlagen bei althochdeutschen Übersetzungstexten ab.

3.6.2 latSyn

Die Ebene *latSyn* enthält den binären Wert *yes/no* und dient der Unterscheidung von Differenz- und Korrespondenzbelegen im Bereich der althochdeutschen Übersetzungstexte (vgl. Kapitel 2.2.4).

Als Abfragebeispiel für *latSyn* sei hier ein Satz mit nicht-lateinischer Syntax gesucht, dessen Mittelfeld eine Verbpartikel enthält:

(25) latSyn="no" & field="Middle" & gf="VPTK" & #1_i_#2 & #2_i_#3

4 DV3DTS – das Tagset

4.1 Allgemeines

Das DV3DTS definiert Tags für die folgenden Kategorien:

- *pos* (Part of Speech)
- *inflection*
- *gf* (grammatical function)
- *is* (information structure)
- *field* (topologische Felder)

im Folgenden wird eine vollständige Liste der in DV3D verwendeten Tags pro Annotationsebene angegeben. Bei *pos* und *inflection* werden hinzugefügte Tags, die keines der DDD-Korpora vorsieht, durch Fettdruck hervorgehoben.

4.2 Tag-Übersicht

4.2.1 pos

Tag	Beschreibung
ADJ	Adjektiv, attributiv
ADJD	Adjektiv, prädikativ
ADJE	Adjektiv, Bestandteil eines Eigennamens
ADJN	Adjektiv, attributiv, nachgestellt
ADJO	Adjektiv, ordinal, attributiv
ADJOS	Adjektiv, ordinal, substituierend
ADJS	Adjektiv, substituierend
ADV	Adverb
ADVNEG	Adverb, negativ
ADVREL	Adverb, relativ
APPO	Postposition
APPR	Präposition / linker Teil einer Zirkumposition
APZR	rechter Teil einer Zirkumposition
CARD	Kardinalzahl, attribuierend
CARDS	Kardinalzahl, attribuierend, substituierend
DD	Determinierer, definit, demonstrativ
DDA	Determinierer, definit, artikelartig
DDD	Determinierer, definit, demonstrativ, prädikativ
DDN	Determinierer, definit, demonstrativ, nachgestellt
DDS	Determinierer, definit, demonstrativ, subst.
DDSREL	Determinierer, definit, substituierend, relativ

DI	Determinierer, indefinit, vorangestellt
DIA	Determinierer, indefinit, artikelartig
DID	Determinierer, indefinit, prädikativ
DIN	Determinierer, indefinit, nachgestellt
DINEG	Determinierer, indefinit, negativ, vorangestellt
DIS	Determinierer, indefinit, substituierend
DPOS	Determinierer, possessiv, vorangestellt
DPOSN	Determinierer, possessiv, nachgestellt
DPOSS	Determinierer, possessiv, substituierend
DWS	Determinierer, interrogativ, substituierend
DWSREL	Determinierer, interrogativ, subst., relativ
ITJ	Interjektion
KOKOM	Konjunktion, vergleichend
KON	Konjunktion, nebenordnend
KOUS	Konjunktion, subordinierend
NA	Nomen Appelativum
NE	Eigenname
NEO	Ortsnamen
PI	Pronomen, indefinit
PINEG	Pronomen, indefinit, negativ
PPER	Personalpronomen
PRF	Pronomen, reflexiv
PTKILL	**Partikel, illokutiv**
PTKINT	Partikel, interrogativ
PTKNEG	Partikel, negativ
PTKREL	Partikel, relativ
PTKVZ	Partikel, Verbzusatz
PTKZU	*zu* vor Infinitiv
PW	Pronomen, interrogativ
PWAV	Pronomen, interrogativ, adverbial
PWAVREL	Pronomen, interrogativ, adverbial, relativ

PWG	Pronomen, interrogativ, generalisierend
PWREL	Pronomen, interrogativ, relativ
VAFIN	Verb, auxiliar, finit
VAIMP	Verb, auxiliar, imperativ
VAINF	Verb, auxiliar, infinitiv
VMFIN	Verb, modal, finit
VMINF	Verb, modal, infinitiv
VMPSN	Verb, modal, PPA, im Verbalkomplex, nachgestellt
VVFIN	Verb, voll, finit
VVIMP	Verb, voll, Imperativ
VVINF	Verb, voll, infinitiv
VVINFS	Verb, voll, infinitiv, substantiviert
VVPP	Verb, voll, Partizip Präteritum, im Verbalkomplex
VVPPA	Verb, voll, Partizip Präteritum, attribuierend
VVPPD	Verb, voll, Partizip Präteritum, prädikativ oder adv.
VVPPN	Verb, voll, Partizip Präteritum, nachgestellt
VVPPS	Verb, voll, Partizip Präteritum, substituierend
VVPS	Verb, voll, Partizip Präsens, im Verbalkomplex
VVPSA	Verb, voll, Partizip Präsens, attributiv
VVPSADV	Verb, voll, Partizip Präsens, adverbial
VVPSD	Verb, voll, Partizip Präsens, prädikativ
VVPSN	Verb, voll, Partizip Präsens, nachgestellt
VVPSS	Verb, voll, Partizip Präsens, substantiviert
$,	Komma
$(	**sonstiges modernes Satzzeichen, satzintern**

4.2.2 INFLECTION

Tag	**Kategorie**	**Beschreibung**
M	Geschlecht	maskulinum
F	Geschlecht	femininum

N	Geschlecht	neutrum
SG	Numerus	singular
PL	Numerus	plural
DU	Numerus	dual
NOM	Kasus	Nominativ
GEN	Kasus	Genitiv
DAT	Kasus	Dativ
ACC	Kasus	Akkusativ
ABL	Kasus	Ablativ
INS	Kasus	Instrumental
IND	Modus	Indikativ
SUBJ	Modus	Konjunktiv
IMP	Modus	Imperativ
PRES	Tempus	Präsens
PRET	Tempus	Präteritum
1	Person	1. Person
2	Person	2. Person
3	Person	3. Person
POS	Komparationsstufe	Positiv
COMP	Komparationsstufe	Komparativ
SUP	Komparationsstufe	Superlativ
WK	Deklination	schwach
ST	Deklination	stark

Das *inflection*-Tier wird im Folgenden Format beschrieben:

Wortart	**Schema**
Substantive	*Genus_Numerus_Kasus*
Pronomen	*Genus_Numerus_Kasus_Deklination*
Adjektive:	*Komparationsst._Genus_Numerus_Kasus_Deklination*

Finite Verben	*Modus_Tempus_Numerus_Person*

4.2.3 GF

Tag	**Beschreibung**
SU	Subjekt
DO	direktes Objekt
IO	indirektes Objekt
GO	Genitivobjekt
PO	Präpositionalobjekt
A	Adverbial
PREDN	Prädikativ
LD	linksversetzte Konstituente
HT	Hanging Topic
FIN	finiter Bestandteil des Prädikats
INF	Infinitiv als Bestandteil einer komplexen Verbform
PART	Partizip als Bestandteil einer komplexen Verbform
VPTK	Verbpartikel
NEG	Negationspartikel des Verbs
ILL	illokutive Partikel (z.B. *eno* und *ia*)
VOC	Vokativ
PAREN	Parenthese

4.2.4 FIELD

Tag	**Beschreibung**
Pre	Vorfeld
Middle	Mittelfeld
Post	Nachfeld
Lsb	Linke Satzklammer

Rsb	Rechte Satzklammer

Tag	Beschreibung
DF	Dislokationsfeld: vergeben für alle Konstituenten, die einen ambigen Status zwischen *Hanging Topic* und *Left Dislocation* aufweisen.
DL	Discourse Linker: vergeben für Partikeln wie *see*, *senu*, sowohl in V3- als auch in VL-Sätzen
IF	illokutives Feld/Illokutionsfeld: vergeben für Illokutionspartikeln wie *inu*, *ja* etc. bei VL-Stellungen

4.2.5 IS

Teiltag	Beschreibung
Frame	temporale oder lokale Domäne, in der eine Proposition verankert wird
Set	Restkategorie
Th	thematisch (vorerwähnter Referent)
Rh	rhematisch (neu eingeführter Referent, nicht vorerwähnt)

Jedes Satzthema wird obligatorisch mit einem Tag-Element $\pm$*sal* erweitert.

Teiltag	Beschreibung
sal	salient (durch den Kontext oder eine Vorerwähnung im Bewusstsein präsenter Referent)

Somit ergeben sich für die *is*-Ebene folgende Tags:

Tag	Beschreibung

Th+sal	alte, bekannte und vorerwähnte Referenten
Th-sal	alte, vorerwähnte Referenten, deren Erwähnung aber schon so weit zurückliegt, dass sie nicht mehr völlig präsent sind
Rh+sal	neue Information, die über die thematischen Informationen nahegelegt wird und aufgrund des Kontextes erwartbar ist
Rh-sal	brandneue Information, auf die der Leser nicht vorbereitet ist, z.B. Einführung völlig neuer Entitäten

5 Literaturverzeichnis

5.1 Editionen

Hench, George A. (Hg.) (1893). *Der althochdeutsche Isidor. Facsimile-Ausgabe des Pariser Codex nebst critischem Texte der Pariser und Monseer Bruchstücke. Mit Einleitung, grammatischer Darstellung und einem ausführlichen Glossar*. Straßburg: Trübner. (= Quellen und Forschungen zur Sprach- und Culturgeschichte der germanischen Völker 72)

Hench, George. A. (Hg.) (1890). *The Monsee Fragments. Newly collated text with introduction, notes, grammatical treatise and exhaustive glossary and a photo-lithographic fac-simile*. Straßburg: Trübner.

Kleiber, Wolfgang (Hg.) (2004). *Otfrid*. Bd. 1: Text. Tübingen: Niemeyer.

Masser, Achim (Hg.) (1994). *Die lateinisch-althochdeutsche Tatianbilingue Stiftsbibliothek St. Gallen Cod. 56*. Göttingen: Vandenhoeck & Ruprecht.

von Steinmeyer, Elias (Hg.) (1971 [1916]). *Die kleineren althochdeutschen Sprachdenkmäler*. Dublin & Zürich: Weidmann.

Tax, Petrus W. (Hg.) (1986-1988). *Die Werke Notkers des Deutschen. Boethius, De consolatione Philosophiæ*. Bd. 1-2. Tübingen: Niemeyer.

Tax, Petrus W. (Hg.) (1979–1983). *Die Werke Notkers des Deutschen. Der Psalter*. Bd. 8–10. Tübingen: Niemeyer.

5.2 Digitale Corpora

Das Bonner Frühneuhochdeutsch-Korpus, Korpora.org (o.D.). URL: `http://www.korpora.org/FnhdC/`.

Donhauser, Karin/Jost Gippert/Rosemarie Lühr (o.D.). ddd-ad (Version 1.0), Humboldt-Universität zu Berlin. Lizenz: CC BY-NC-SA 3.0. url: https://www.deutschdiachrondigital.de.

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

5.3 Corpus-Tools

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

Schmidt, Thomas/Kai Wörner (2014). EXMARaLDA. In: *Handbook on Corpus Phonology*. Jacques Durand/Ulrike Gut/Gjert Kristoffersen (Hgg.). Straßburg: Oxford University Press, 402–419. url: https://exmaralda.org.

Zipser, Florian/Laurent Romary (2010). A model oriented approach to the mapping of annotation formats using standards. Workshop on Language Resource and Language Technology Standards, LREC 2010, May 2010, La Valette, Malta. inria-00527799. url: https://hal.archives-ouvertes.fr/inria-00527799/en/.

5.4 Tagsets

Dipper, Stefanie/Karin Donhauser/Thomas Klein/Sonja Linde/Stefan Müller/Klaus-Peter Wegera (2013). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Com-*

putational Linguistics 28: 1. Special Issue, 85–137. URL: https://jlcl.org/content/2-allissues/9-Heft1-2013/5Dipper.pdf.

Schiller, Anne/Simone Teufel/Christine Stöckert/Christine Thielen (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technischer Bericht. URL: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf, Stuttgart & Tübingen.

5.5 Sekundärliteratur

Axel, Katrin (2007). *Studies on Old High German Syntax. Left Sentence Periphery, Verb Placement and Verb Second.* Amsterdam: Benjamins.

Bader, Markus/Tanja Schmid (2009). Verb clusters in colloquial German. *The Journal of Comparative Germanic Linguistics* 12, 175–228.

Behagel, Otto (1932). *Deutsche Syntax : eine geschichtliche Darstellung. Bd. IV. Wortstellung. Periodenbau.* Heidelberg: Winter.

Chafe, Wallace L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In: *Subject and Topic.* Charles N. Li (Hg.). New York: Academic Press, 25–55.

Demske, Ulrike (2008). Syntax and discourse structure : verb-final main clauses in German. In: *Non-Canonical Verb Positioning in Main Clauses. Special Issue of Linguistische Berichte 25.* Mailin Antomo/Sonja Müller (Hgg.), 137–159.

Fleischer, Jürg (2006). Zur Methodologie althochdeutscher Syntaxforschung. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 128, 25–69.

Klein, Thomas/Stefanie Dipper (2016). Handbuch vom Referenzhandbuch Mittelhochdeutsch. *Linguistische Arbeitsberichte* 19.

Krifka, Manfred (2008). Basic Notions of Information Structure. *Acta Linguistica Hungarica* 55, 243–276.

Masser, Achim (1991). *Die lateinisch-althochdeutsche Tatianbilingue des Cod. Sang. 56. Mit zwölf Abbildungen.* Göttingen: Vandenhoeck und Ruprecht. (= Nachrichten der Akademie der Wissenschaften in Göttingen I. Philologisch-historische Klasse, Nr. 3)

Molnár, Valeria (1993). Zur Pragmatik und Grammatik des TOPIK-Begriffs. In: *Wortstellung und Informationsstruktur.* Marga Reis (Hg.). Tübingen: Niemeyer, 155–202.

Petrova, Svetlana/Michael Solf (2009). On the Methods of Information-Structural Analysis in Texts from Historical Corpora : A Case Study on Old High German. In: *Information Structure and Language Change : New Approaches to Word Order Variation in Germanic.* Roland Hinterhölzl/Svetlana Petrova (Hgg.). Berlin: de Gruyter, 121–160.

Winkler, Julia (2017). V3-Stellung im Deutschen : Wettbewerb um das Vorfeld. *Linguistische Berichte* 250, 139–168.