Annotationsguidelines zur Deutschen Diachronen Baumbank (DDB)

Projektinfo: http://korpling.german.hu-berlin.de/ddb-doku/index.htm
Suchinterface: http://korpling.german.hu-berlin.de/ddd/
('DeutscheDiachroneBaumbank' auswählen)

Zitieren Sie das Dokument bitte wie folgt:
Hirschmann, Hagen; Linde, Sonja (2010): Annotationsguidelines zur Deutschen Diachronen Baumbank. Technical Report. Humboldt-Universität zu Berlin.

Das durch den Berliner Senat geförderte Projekt "Interdisziplinärer Forschungsverbund Linguistik - Bioinformatik zur Berechnung von Verwandtschaft und Abstammung" hat angestrebt, Wege zu finden, wie bioinformatische Methoden dazu verwendet werden können, die Verwandtschaft zwischen (schriftlichen) Sprachdaten automatisch messbar zu machen.

Um dies verwirklichen zu können, braucht es miteinander verwandte Texte, die bis auf die Tatsache, dass sie in unterschiedlichen Zeiten bzw. Sprachstufen entstanden sind, möglichst vergleichbar sind (z.B. weil sie Texte derselben Textsorte sind oder aus demselben Sprachraum stammen) und die auf dieselbe Art und Weise annotiert wurden. In einem hier vorzustellenden Teilprojekt wurden deshalb exemplarisch möglichst authentische Texte aus drei verschiedenen Sprachstufen des Deutschen (Althochdeutsch, Mittelhochdeutsch und Frühneuhochdeutsch) auf unterschiedlichen grammatischen Ebenen annotiert. Am wesentlichsten hinsichtlich des Projektvorhabens ist die Ebene der syntaktischen Annotation, denn das Vorhaben zielt in erster Linie auf die Messung syntaktischer Unterschiede zwischen Sprachen ab. Annotationsebenen unterhalb der syntaktischen Ebene können insbesondere dazu dienen, bestimmte syntaktische Elemente zu identifizieren bzw. grammatische Klassen genauer zu differenzieren. Die Annotation geschah so nah wie möglich an Prinzipien, die sich für das Neuhochdeutsche bewährt haben, denn letztendlich sollen die älteren Sprachstufen auch mit neuhochdeutschen Korpora vergleichbar sein. Da das Tiger-Korpus[1] mit ca. 1000000 Token die derzeit größte syntaktische Baumbank des Deutschen ist, die voraussichtlich die beste Grundlage für Berechnungen von syntaktischen Unterschieden bietet, und weil zwei gleich annotierte Datenmengen am geeignetsten für die Messung sprachlicher Unterschiede sind, wurde sich bei der Entwicklung des Annotationsschemas für die drei älteren Sprachstufen des Deutschen so weit wie möglich an den bestehenden Annotationsrichtlinien, die im Rahmen des Tiger-Projektes erstellt wurden, orientiert.

Im Folgenden werden die grundlegenden Entscheidungen des Teilprojekts dargestellt und begründet sowie die unterschiedlichen Annotationsebenen vorgestellt. Da auf den Ebenen der

[1] Das Tiger-Korpus wurde zwischen 1999 und 2005 an dem Insitut für Computerlinguistik an der Universität des Saarlandes, am Insitut für Maschinelle Sprachverarbeitung der Universität Stuttgart sowie am Insitut für Germanistik der Universität Potsdam erstellt. Für nähere Informationen siehe http://www.ims.uni-stuttgart.de/projekte/TIGER/.

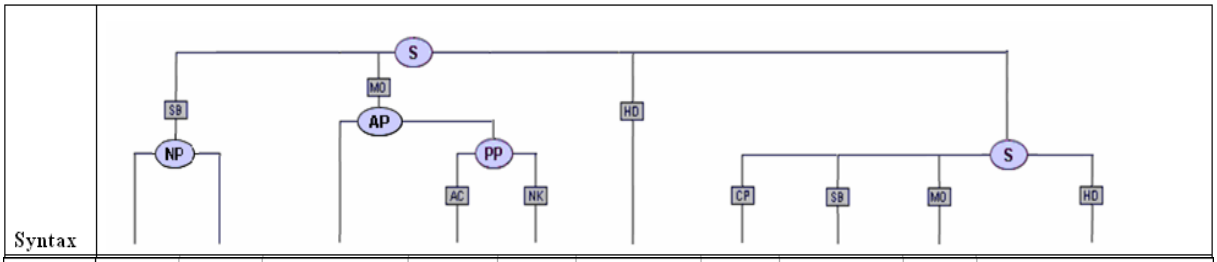
Lemmata, der Wortarten, der Morphologie und der Syntax prinzipiell nach bereits existierenden Richtlinien vorgegangen wurde, werden hier lediglich die Abweichungen bzw. Ergänzungen von den zugrunde liegenden Schemata dargestellt. Dies erfolgt für die drei unterschiedlichen Sprachstufen separat (es gibt z.T. abweichende Konventionen zwischen den einzelnen Sprachstufen). Zudem werden Beispiele für die korpusbasierte Suche nach bestimmten, z.T. sprachspezifischen grammatischen Kategorien und Klassen gegeben, um den potentiellen NutzerInnen der entstandenen Subkorpora die Korpusabfrage möglichst leicht zu gestalten.

INHALT

1. Allgemeine Annotationsprinzipien (Ebenen, die für alle drei Subkorpora gelten):

Die Sprachdaten sind auf unterschiedlichen linguistischen Ebenen annotiert. Diese Annotationen sind in einer Mehrebenenarchitektur gespeichert, wodurch zum einen die einzelnen Ebenen unabhängig voneinander durchsucht werden können und zum anderen zueinander in Beziehung gesetzt werden können.

In der folgenden Abbildung sind die grammatischen Annotationen zu sehen, die für alle Subkorpora gelten:



Normal.	zuene	plinte	sizsente		bi	uuege	gahortun	daz	iesus	dar	fuor
Hench Edit.	zuene	*plinte*	siz	/ cente	biuuege		ga hortun	daz	ihs	dar	fuor
Lemma	zwene	blint	sizzen		bi	weg	gihoren	thaz	iesus	thar	faran
POS	CARD	NN	ADJD		APPR	NN	VVFIN	KOUS	NE	ADV	VVFIN
Morph.	Nom.Pl.*	Nom.Pl.*	Nom.Pl.*			Dat.Sg. Masc	3.Pl.Past.Ind		Nom.Sg.Masc		3.Sg.Past.Ind

Satz aus dem DDB-Althochdeutschkorpus mit Schriftnormalisierung, Editionstranskription, Lemmatisierung, Wortartenwerten und morphologischen Werten

http://korpling.german.hu-berlin.de/ddd/Cite/AQL(%22gahortun%22),CIDS(DDB.AHD),CLEFT(5),CRIGHT(4)

Als weitere Annotationsebenen kommen für alle Subkorpora bibliographische Angaben hinzu und für den althochdeutschen Text wurden die lateinischen Übersetzungsvorlagen in das Korpus mit aufgenommen (s.u.).

Der Syntaxbaum stellt die syntaktische Annotation dar. Ihre Prinzipien bestehen im Wesentlichen aus der Darstellung von Phrasenstrukturen und syntaktischen Abhängigkeiten.

Die Einheiten auf der 'word'-Ebene (Tokens bzw. Wörter) werden in einem übergeordneten Knoten zu Phrasen bzw. Konstituenten zusammengefügt. Der Knoten erhält dabei ein der syntaktischen Kategorie entsprechendes Label. Phrasen können Bestandteile von weiteren Phrasen sein; dementsprechend werden die unteren Knoten mit darüber liegenden verbunden. Die Abhängigkeiten der einzelnen Wörter und Phrasen und vor allem ihre Funktion im Satz werden durch Kantenlabels ausgedrückt. Sämtliche Elemente, die zu einem Satz gehören, laufen letztendlich in einem Satzknoten (mit dem Label 'S') zusammen. Zwischen einzelnen Sätzen werden keine syntaktischen Relationen angezeigt bzw. geht die syntaktische Annotation nicht über Satzgrenzen hinaus.

Der morphologischen, syntaktischen sowie der Wortartenannotation in diesem Projekt liegt das Annotationsschema[2] des TIGER-Projektes[3] zugrunde, um eine möglichst gute Vergleichbarkeit zu dem größten syntaktisch annotierten Korpus des Neuhochdeutschen zu gewährleisten. An den dort vorgestellten Tagsets für Knoten- und Kantenbezeichnungen sowie Wortarten wurden keine Änderungen vorgenommen, lediglich im Bereich der flexionsmorphologischen Analyse mussten ein paar wenige Kategorien hinzugefügt werden. Lediglich die Annotationsprinzipien bei bestimmten sprachlichen Strukturen/Konstruktionen wurden modifiziert oder erweitert, sofern sie mithilfe der bestehenden Richtlinien nicht hätten

[2] Erhältlich unter www.ifi.uzh.ch/CL/volk/treebank_course/**tiger**_annot.pdf (Zugriff am 15.08.2008)

[3] Vgl. z.B. http://www.ims.uni-stuttgart.de/projekte/TIGER/ (Zugriff am 15.08.2008)

beschrieben oder suchbar gemacht werden können. Solche Änderungen werden weiter unten für die einzelnen Sprachstufen bzw. Subkorpora getrennt aufgeführt.

1.1 Erstellung einer normalisierten Wortformebene

Die unterschiedlichen Texteditionen sind sehr uneinheitlich bezüglich vieler Faktoren, die mit der Repräsentation letztendlich der gesprochenen Sprache zusammenhängen. Die nennenswertesten Probleme hierbei sind die uneinheitliche Spatiensetzung (in älteren Sprachstufen des Deutschen besteht mitunter noch keine Konvention hinsichtlich dessen, was als (graphematisches) Wort angesehen wird), die uneinheitlichen Graphemsysteme (bspw. finden sich in manchen Editionen bestimmte Diakritika für die Explizierung phonologischer Gegebenheiten) oder unterschiedliche Abkürungskonventionen.

Um die Suche über die unterschiedlichen Texte hinweg überhaupt möglich und auch möglichst intuitiv zu machen, wurde die Editionsebene, welche sämtliche Informationen der Textedition enthält, mit einer normalisierten Ebene alligniert.

Normal.	zuene	plinte	sizsente			bi	uuege	gahortun		daz	iesus	dar	fuor			
Hench Edit.	zuene	*plinte*	siz	/	cente	biuuege		ga	hortun	daz	ihs	dar	fuor			

Beispiel für die Normalisierung von Wortformen im althochdeutschen Subkorpus: Es werden Wortformen aufgelöst (biuuege --> bi uuege), zusammengeführt (ga hortun --> gahortun), Zeilentrennungen aufgehoben (siz / cente --> sizsente) , Kursivierung aufgelöst (*plinte* --> plinte), Abkürzungen aufgelöst (ihs --> iesus),.

Da sich die Normalisierung nach der Beschaffenheit der jeweiligen Edition richtet, werden die unterschiedlichen Normalisierungen im textspezifischen Teil (unten) behandelt.

1.2 Lemmatisierung

Um nach unterschiedlichen Wortformen, die zu derselben Grundform gehören, suchen zu können, wird jeder Wortform (auf der normalisierten Ebene) ein Lemma (=eine Grundform) zugeordnet. Dieses ist (per Konvention) bei Verben der Infinitiv und bei den nominalen Einheiten die erste Person Singular (ggf. Maskulinum, stark).

Bei diesem Schritt findet zudem eine orthographische Normalisierung nach einer für die entsprechende Sprachstufe geltenden Norm statt, die durch ein entsprechendes Referenzwörterbuch gegeben wird. Somit ist es auf den unterschiedlichen Ebenen möglich, entweder nach Formen zu suchen, welche denjenigen der uns vorliegenden Textedition entsprechen (dies ist auf der entsprechenden Editionsebene möglich). Möchte man jedoch etwas intuitiver nach Wortformen ohne editionsspezifische Sonderzeichen oder Schreibungen suchen, so kann man auf der Lemma-Ebene nach Grundformen suchen, die den Einträgen des für die jeweilige Sprachstufe gültigen Referenzwörterbuchs entsprechen.

Normal.	zuene	plinte	sizsente	bi	uuege	gahortun	daz	iesus	dar	fuor
Hench Edit.	zuene	*plinte*	siz / cente	biuuege		ga hortun	daz	ihs	dar	fuor
Lemma	zwene	blint	sizzen	bi	weg	gihoren	thaz	iesus	thar	faran

Beispiel für die Lemmatisierung von Wortformen im althochdeutschen Subkorpus: Zum einen werden Wortformen auf eine Grundform zurückgeführt (plinte --> plint), zum anderen werden individuelle orthographische Repräsentationen gemäß Wörterbuch (hier: Schützeichel 1995) normalisiert (sizcen --> sizzen)

1.3 Wortartentagging

Jeder Wortform auf der Normalisierungsebene wird ein Tag für eine Wortart zugeordnet. Das verwendete Tagset ist dasjenige, welches sich im Deutschen am besten durchgesetzt hat bzw. für die meisten Korpora verwendet wird – das STTS (Stuttgart-Tübingen Tagset)[4]. Die vorgenommenen Änderungen an dieser Liste entspricht denen des Tiger-Projekts, um mit den in diesem Projekt annotierten Daten Vergleichbarkeit zu gewährleisten.[5]

Das STTS differenziert bei einigen flektierbaren Wortarten die Vorkommen nach morphosyntaktischen Merkmalen (wie Finitheit) aus, im pronominalen Bereich werden die Vorkommen nach den Kriterien attribuierend und substituierend getrennt. Somit ergeben sich Redundanzen bzw. inhaltliche Überschneidungen der Wortartenannotation mit sowohl der morphologischen als auch der syntaktischen Annotation. Dies muss allerdings kein Nachteil für die Suche in den Daten sein; vielmehr können Elemente so auf unterschiedliche Weise gefunden werden, was gerade für weniger geübte KorpusnutzerInnen eine Erleichterung bedeuten kann.

Normal.	zuene	plinte	sizsente	bi	uuege	gahortun	daz	iesus	dar	fuor
Hench Edit.	zuene	*plinte*	*siz* / sente	biuuege		ga hortun	daz	ihs	dar	fuor
Lemma	zwene	blint	sizzen	bi	weg	gihoren	thaz	iesus	thar	faran
POS	CARD	NN	ADJD	APPR	NN	VVFIN	KOUS	NE	ADV	VVFIN

Beispiel für die Zuordnung von Wortartentags zu den normalisierten Wortformen im althochdeutschen Subkorpus

1.4 Morphologisches Tagging

Zu einer umfassenden syntaktischen Analyse von stark flektierenden Sprachen wie den hier beschriebenen Sprachstufen des Deutschen gehört die flexionsmorphologische Beschreibung der im Satz enthaltenen Wortformen. Die Berücksichtigung dieser Annotationsebene kann bei der Suche nach bestimmten grammatischen Kategorien von erheblichem Nutzen sein.

Die Richtlinien für das flexionsmorphologische Tagging sind weitestgehend angelehnt an die Richtlinien für neuhochdeutsche Zeitungstexte aus dem Tiger-Projekt (Crysmann et al. 2005)[6].

[4] Eine Einführung in das STTS sowie die Liste von Wortartentags und deren Verwendung befindet sich unter der Internetadresse http://www.sfb441.uni-tuebingen.de/a5/codii/info-stts-de.xhtml

[5] Namentlich betrifft dies erstens die Unterscheidung PIDAT/PIAT – alle attribuierenden Indefinitpronomen werden als PIAT getaggt; zweitens heißt das beim STTS vorgesehene Tag PAV PROAV (die Verwendung ist dieselbe).

6 Beziehbar unter https://files.ifi.uzh.ch/cl/siclemat/lehre/papers/tiger-morph.pdf

Hierbei wird den flektierbaren Wortarten ein Wert gemäß ihrem aktuellen Flexionsstatus im Satz zugewiesen. Generell nicht flektierbare Wortarten erhalten keine Annotation ("--").

Normal.	zuene	plinte	sizsente					bi	uuege	gahortun		daz	iesus	dar	fuor									
Hench Edit.	zuene	*plinte*	*siz*	/	sente	•	biuuege			ga	hortun	daz	ihs	dar	fuor									
Lemma	zwene	blint	sizzen					bi	weg	gihoren		thaz	iesus	thar	faran									
POS	CARD	NN	ADJD					APPR	NN	VVFIN		KOUS	NE	ADV	VVFIN									
Morph.	Nom.Pl. Masc	Nom.Pl. Masc	Nom.Pl. Masc						Dat.Sg. Masc	3.Pl.Past.Ind			Nom.Sg.Masc		3.Sg.Past.Ind									

Beispiel für die morphologische Annotation der normalisierten Wortformen im althochdeutschen Subkorpus

Wörter, die (syntaktischen) Wortarten angehören, welche prinzipiell flektierbar sind, die aber allgemein oder in einem bestimmten Kontext unflektierbar sind, erhalten für die nicht zu markierende Kategorie den Wert "*". (Am häufigsten betrifft dies die Wortart Pronomen, in der es sowohl in attributiver als auch in substituierender Verwendung unflektierbare Wörter gibt, z.B. *allerlei* oder *manch*.[7]) Manchmal betrifft dies nur bestimmte Kategorien, d.h. gewisse flexionsmorphologische Kategorien können markiert werden. Dann erhalten nur diese jeweiligen Kategorien den Wert "*".[8]

Der Wert "*" dient nicht nur zur Markierung von Unflektierbarkeit, sondern auch zur Kennzeichnung von flexionsmorphologischen Ambiguitäten bei Synkretismen. Im Falle, dass sich über den syntaktischen Kontext nicht klar eine von mindestens zwei synkreten Formen desambiguieren lässt, wird auch der Wert "*" für die ambige(n) Kategorie(n) vergeben.[9]

Lediglich auf dieser Annotationsebene wurden – entgegen den anderen an das Tiger-Projekt angelehnten Annotationsebenen – Änderungen hinsichtlich der vorliegenden Richtlinien vorgenommen. Diese Änderungen hängen von der jeweiligen Sprachstufe ab und werden dementsprechend bei den einzelnen Subkorpora besprochen.

1.5 Syntaxannotation

Die syntaktische Annotation erfolgt durch dir Erstellung von Baumgraphen, die Knoten und Kanten enthalten. Sie ist weitestgehend angelehnt an das TIGER-Annotationsschema[10]. Die Knoten entsprechen syntaktischen Phrasenkategorien, die Kanten syntaktischen Funktionen und Bindungsrelationen. Somit verbindet die syntaktische Annotation eine Phrasenstrukturbeschreibung (mit Kategorien wie "AP" oder "NP" für "Adjektiv-" bzw. "Nominalphrase" bzw. Kantenlabels wie "HD" für "Kopf") mit einer dependenzgrammatischen Beschreibung (z.B. durch das Kantenlabel "OA" für "Objekt, Akkusativ").

Das Annotationsschema kann erheblich von aktuellen generativen oder dependenziellen Syntaxformaten abweichen. Es erhebt keinen Anspruch auf eine adäquate modellgebundene

7 Uns ist bewusst, dass diese Wortarteinteilung streitbar bis linguistisch fragwürdig ist. Sie geht zurück auf die Klassifikation des STTS-Tagsets.

8 Bspw. kann bei Personalpronomina bis auf bei der 3. Person Singular (er, sie, es) nicht bestimmt werden, um welches Genus es sich handelt.

9 Im Neuhochdeutschen entstehen z.B. viele Genitiv-Dativ-Ambiguitäten nach Präpositionen, die bezüglich dieser Kasus uneinheitlich funktionieren (*trotz (?Gen/Dat)seiner (?Gen/Dat)Vergangenheit*).

10 Beziehbar unter http://www.linguistics.ruhr-uni-bochum.de/~dipper/papers/tiger_annot.pdf

Beschreibung, sondern soll bei der Suche nach unterschiedlichen syntaktischen Konstruktionen möglichst einheitlich und funktional sein.

Bei der syntaktischen Annotation der verschiedenen Sprachstufen des Deutschen wurden hinsichtlich dieser Vorlage keine Änderungen an den Listen für Knoten- und Kantenlabels vorgenommen. Dies ermöglicht vor allem eine (sub-)korpusübergreifende Suche in vier Sprachstufen mit einheitlichen Kategorie- und Funktionsbezeichnungen.

Dies hatte zur Folge, dass bestimmte syntaktische Kategorien, Funktionen oder Relationen, spezifisch für eine bestimmte Sprachstufe, nicht explizit (durch Vergabe von entsprechenden Knoten- oder Kantenlabels) gelabelt wurden, sondern implizit durch distinktive Annotationskonventionen beschrieben werden. Beispielsweise kann dies durch eine (ansonsten nicht den Konventionen entsprechende) Verknüpfung von morphologischer und syntaktischer Annotation geschehen.

Die für die einzelnen Sprachstufen geltenden Sonderkonventionen werden in den folgenden Abschnitten beschrieben.

2. Subkorpora der einzelnen Sprachstufen AHD, MHD, FNHD:

Obwohl die Annotationsrichtlinien, wie im vorigen Kapitel 1 beschrieben, für die drei Sprachstufen möglichst einheitlich sein sollen, stellt doch jeder einzelne bearbeitete Text eigene Anforderungen und muss bezüglich bestimmter Merkmale gesondert behandelt werden. Dies liegt vor allem an den höchst individuellen Beschaffenheiten, wie die alten Texte erhalten bzw. überliefert sind. Es wird demnach sowohl auf die allgemeinen annotationsprinzipien, die für alle Texte/Sprachstufen gelten, als auch auf die für den jeweiligen Text/die jeweilige Sprachstufe spezifischen Annotationsprinzipien eingegangen.

Die in das Modellkorpus aufgenommenen Texte wurden unter der Zielsetzung der bestmöglichen Vergleichbarkeit ausgewählt. Sie entstammen einem Dialektgebiet – dem Bairischen, wenn auch an manchen Stellen Einsprengsel anderer Mundarten auftreten können – und gleichen sich in der Textsorte (sie sind alle religiöse Texte). Das (eigentlich selbstverständliche Ziel), möglichst authentische und autochthone Texte nicht zu geringen Umfangs auszuwählen, ist für die älteste Sprachstufe des Deutschen, dem Althochdeutschen, nur schwer zu realisieren. Der Großteil der überlieferten ahd. Texte sind Übersetzungstexte, die sich i.A. ziemlich strikt an eine zumeist lateinische Vorgabe halten. Der einzige größere Text, der nicht direkt dem Lateinischen entstammt – Otfrid von Weissenburgs *Evangelienharmonie* – ist im Endreim verfasst und ist wegen des metrischen Zwanges besonders für die syntaktische Analyse wenig geeignet. Ein Korpus aus mehreren kleineren Texten zusammenzustellen, ist ebenfalls wenig attraktiv, da dieses Korpus in allen Bereichen (z.B. Dialekt, Alter, Textsorte, usw.) ausgesprochen heterogen wäre. Zudem sind längere zusammenhängende Passsagen wünschenswert, um auch textlinguistische Untersuchungen durchführen zu können. Viele Überlieferungen sind nur so fragmentarisch erhalten, viele Satzeinheiten unterbrochen sind und sich deshalb eine syntaktische Analyse ausschließt.

Aus diesen Gründen wurden für das ahd. Subkorpus die sogenannten *Monseer Fragmente* ausgewählt, einer der ältesten volkssprachlichen Texte überhaupt. Ihr hohes Alter machen die *Monseer Fragmente* gegenüber anderen, jüngeren ahd. Übersetzungstexten in besonderem Maße interessant.

Für die Auswahl der Texte in den anderen Sprachstufen wurde darauf geachtet, dass sie möglichst vergleichbar hinsichtlich Textsorte und Sprache (Dialektgebiet) sind. Da es mit fortschreitender Zeit stetig mehr (und besser erhaltene) Texte gibt, war an dieser Stelle die Suche nach geeignetem Material nicht so problematisch. Um jedoch für das

Mittelhochdeutsche und das Frühneuhochdeutsche Übersetzungsphänomene, welche die Authentizität des Textes mindern, ausschließen zu können, wurden für diese beiden Sprachstufe keine Übersetzungen des Matthäus – Evangeliums ausgewählt, sondern Predigttexte, die zwar ebenfalls auf Vorlagen zurückgehen, welche jedoch keine engen Übersetzungen sind. Für das mhd. Subkorpus wurden die so genannten *Speculum Ecclesiae deutsch* und für das frnhd. Subkorpus Predigten des Veit Nuber bearbeitet.

Bei der Annotation stand der Anwendungsnutzen im Vordergrund, d.h. die Vergleichbarkeit der bearbeiteten Subkorpora, so dass das diachrone Gesamtkorpus möglichst einheitlich annotiert sein sollte. Dabei diente das Tiger-Annotationsschema für neuhochdeutsche Zeitungstexte die Grundlage für die älteren Sprachstufen. Bereits dieses Annotationsschema erhebt keinen Anspruch auf eine korrekte grammatische Beschreibung der syntaktischen Strukturen, sondern möchte hinsichtlich der Suche syntaktischer Strukturen möglichst funktional sein. Bestimmte Annotationsprinzipien auch in diesem Projekt können also theoretisch fraglich erscheinen; sie sollen jedoch keine adäquate linguistische Beschreibung illustrieren, sondern gehen hervor aus dem Bemühen um Vergleichbarkeit einzelner grammatischer Strukturen in den verschiedenen Subkorpora/Sprachstufen des Deutschen und der Möglichkeit ihres systematischen Vergleichs bzw. ihrer Suche im Gesamtkorpus.

Am Ende werden einzelne Zweifelsfälle besprochen.

2.1 Richtlinien für die Annotation des Althochdeutschen

Die Annotation des althochdeutschen Korpus gliedert sich in folgende Ebenen:

- [Hench], Textwiedergabe nach der Edition
- [Word], Bearbeiteter Text
- [Lat], Lateinische Referenzebene
- [Bibl], Bibliographische Referenzebene
- [Lemma], Lemmatisierung der Wortformen
- [POS], Zuweisung von Wortart an die einzelne Wortform
- [Morph], Kennzeichnung morphologischer Merkmale einer Wortform
- [Syntax], Syntaktische Kommentierung

2.1.1 Text

Grundlage des althochdeutschen Korpus ist das Matthäus-Evangelium der Mon(d)seer Fragmente (auch: Monsee-Wiener Fragmente, Fragmenta theodisca; ÖNB cod. 3093)

Bei den Mon(d)seer Fragmenten handelt es sich um Bruchstücke verschiedener religiöser Texte in bairischer Sprache, die wahrscheinlich einer Handschrift entstammen und der Isidor-Sippe zugehörig sind. Die Handschrift wird auf das Ende des 8. Jahrhunderts datiert und wurde vermutlich in dem Kloster Monsee (heute Mondsee) abgefasst. Das in diesen Fragmenten enthaltene Matthäus-Evangelium ist die älteste deutsche Übersetzung eines der vier Evangelien.

Edition: Hench, George Allison, 1890. The Monsee Fragments. Strassburg: Karl J. Trübner.

2.1.2 Edition und Textwiedergabe

Bei der verwendeten Edition handelt es sich um eine quasi-diplomatische, d.h. sie folgt der Handschrift in der Wiedergabe der Schreibweise, Interpunktion, Initialen, Abkürzungen und Spatien zwischen morphographischen oder lexikalischen Einheiten. Auch wird der Text zeilengetreu wiedergegeben, die Zeilen sind nummeriert. Neben den deutschen Text wird der lateinische, der im Original jeweils auf der Rückseite des deutschen Textes erscheint, zeilengenau gestellt. Ein reichhaltiger Appendix weist auf Unsicherheiten in der Lesart, Radierungen und Korrekturen, Beschädigungen des Pergaments, Lücken usw. hin. Modernen Ansprüchen an eine handschriftengerechte Textwiedergabe genügt die vorliegende Ausgabe allerdings nur bedingt, da der Herausgeber in verschiedenen Punkten in den Text eingreift. So werden „evident scribal errors“ (Hench 1890:XXV) vom Herausgeber korrigiert und ausgelassene Buchstaben und Wörter eingefügt, ebenso wie bei fragmentarischen Textstellen Buchstaben oder gar ganze Wörter erschlossen und in den Text integriert werden. Sämtliche Änderungen sind entweder, wie die Korrekturen des Herausgebers, im Anhang vermerkt oder direkt im Text gekennzeichnet; Einfügungen werden durch Klammerung und „Rekonstruktionen“ durch Kursivschreibung markiert.

Der Text wurde getreu der Ausgabe fortlaufend unter Beibehaltung der Kennzeichnung der Änderungen, der Sonderzeichen <æ>, <ħ>, <ū> und der Interpunktion digitalisiert. Asterices, die in schwankender Anzahl die Größe einer Lücke illustrieren sollen, werden auf drei <***> gekürzt. Da aus technischen Gründen kein ‚s’ und kein ‚p’ mit hochgestelltem Strich dargestellt werden können, werden diese als <s̄> bzw. als <p̄> umschrieben.

2.1.3 Normalisierte Wortebene

Wortformen wurden vereinheitlicht, indem getrennt geschriebene lexikalische Einheiten zusammengefügt wurden und umgekehrt mehrere, zusammengeschriebene Lexeme getrennt wurden. Um einfach nach Wortabfolgen suchen zu können, wurde die Interpunktion auf dieser Ebene aufgehoben. Des Weiteren wurden die Klammerung und Kursivschreibung aufgelöst, welche die Suche nach Wortformen ebenso behindern würden. Orthographische Konventionen wurden auf dieser Ebene beibehalten.

Da der Text zu Teilen fragmentarisch ist, wurden nur syntaktisch vollständige Sätze bearbeitet. Unvollständige Strukturen würden falsche Suchergebnisse liefern und Statistiken verfälschen.

2.1.4 Lateinische Referenzebene

Der lateinische Text der Ausgabe wurde zeilengenau übernommen, so dass Übereinstimmungen und Abweichungen zwischen lateinischem Original und deutscher Übersetzung unmittelbar im Korpus überprüft werden können.

2.1.5 Bibliographische Referenzebene

Die bibliographischen Angaben beziehen sich jeweils auf eine Zeile in der verwendeten Ausgabe. Sie bezeichnen 1. den Herausgeber der Ausgabe, 2. die Seite in der Ausgabe, 3. die Abbreviatur des Textes mit der Kapitelnummer und 4. die Zeilenangabe, also z.B. *Hench/11/M-VIII/4*.

2.1.6 Lemmatisierung der Wortformen

Auf der Lemmatisierungsebene wird nicht nur jeder Wortform eine Grundform zugewiesen, sondern aufgrund individueller und inkonsistenter Schreibweisen eine orthographische Normalisierung durchgeführt. Im althochdeutschen Korpus wird jeweils die Form, die das etablierte Wörterbuch Schützeichels (Schützeichel 1995) ausweist, übernommen. Es wird das jeweils erste Lexem eines Artikels als Lemma verarbeitet, wobei dieses im Allgemeinen der Form des althochdeutschen Tatians entspricht. Sofern eine Form nicht im Wörterbuch belegt ist, wird die Orthographie der Edition übernommen.

Aufgrund bestimmter syntaktischer Annahmen (s.u.) weichen wir in wenigen Fällen vom Lexikoneintrag des Wörterbuchs ab, die Abweichungen werden in der folgenden Liste aufgeführt.

Wörterbuch	Abweichung
aer danne	aerdanne
after diu	afterdiu
diu mera	diumera
wela tuoan	welatuoan

Abweichungen von Wörterbucheintrag und Lemmaeintrag

2.1.7 Zuweisung der Wortart

Im Folgenden werden Zweifelsfallentscheidungen bei der Tagzuweisung im ahd. Korpus aufgeführt.

Definitartikel / Demonstrativpronomen

Das Althochdeutsche ist als Artikelsprache umstritten. Im Allgemeinen wird angenommen, dass sich der bestimmte Artikel sukzessive aus dem formgleichen Demonstrativpronomen entwickelt hat. Da keine Einheitlichkeit im Gebrauch des Artikels besteht und in vielen Fällen semantisch nicht zwischen dem Status Artikel und Demonstrativpronomen unterschieden werden kann, werden im ahd. Korpus alle Artikelwörter, die auf das Lemma „ther“ zurückgeführt werden können, als [PDAT] getaggt.

Indefinitartikel / Numerale / Adjektiv / Indefinitpronomen

Der Gebrauch des Indefinitartikels ("*ein*") im Ahd. ist schwer zu bestimmen. Wenn man eine solche Wortklasse annimmt, so scheint "*ein*" dennoch vor allem in spezifischer Lesart aufzutreten. „*ein*“ flektiert adjektivisch, es treten sowohl die schwache als auch die starke Flexion auf. Außer als Zahlwort tritt es auch als

Indefinitpronomen ‚irgendeiner' und als Adjektiv ‚einziger' auf. Oftmals fällt es schwer, die präzise Bedeutung und Funktion von „ein" zu ermitteln.

Numerale / Adjektiv

Die Kardinalzahlen von ‚eins' bis ‚zwölf' flektieren im Ahd. oftmals. Wir vergeben für alle Zahlen das Tag [CARD], auch wenn sich diese im Einzelfall wie Adjektive verhalten.

Verb / prädikatives Adjektiv

Der Status der Partizipien im Ahd. ist umstritten. Sowohl das Partizip Präsens als auch das Partizip Präteritum treten zum Bezugsnomen kongruent flektiert in prädikativer Position oder als Teil einer verbalen Fügung auf, durchaus aber ebenso unflektiert in den gleichen Funktionen. Die Bedeutung und die Funktion der mehrgliedrigen Verbalformen mit Partizip können häufig nicht eindeutig festgestellt werden und werden in der Forschung kontrovers diskutiert, zumal dass nhd. Verbalsystem mit seinen relativ eindeutigen Periphrasen und dem darin integrierten Partizip Präteritum sicherlich nicht als Schablone für die ahd. Fügungen dienen kann. In den einschlägigen Grammatiken werden die Partizipien mit dem Infinitiv Präsens als „Verbalnomina" zusammengefasst. Die Zuweisung einer entsprechenden Kategorie an die Partizipien und den Infinitiv hätte gewiss den Vorteil, dass die fraglichen Konstruktionen ohne die Implikation einer bestimmten theoretischen Interpretation abbildbar wären und innerhalb des synchronen Subkorpus die Verbalnomina eindeutig erfragt werden könnten, vor dem Hintergrund der diachronen Anlage des Gesamtkorpus und der damit verbundenen Forderung nach der Vergleichbarkeit der möglichen Abfragen und Suchergebnisse erscheint jedoch ein solches Verfahren als wenig praktikabel. Aus diesem Grund haben wir uns dafür entschieden, bei der Annotation der Partizipien den Konventionen des Nhd. zu folgen und das Partizip Präsens, wenn es als zweiter, infiniter Bestandteil einer mehrgliedrigen Verbalform, also in prädikativer Stellung im weiteren Sinne, als [ADJD], das Partizip Präteritum in entsprechender Stellung jedoch als [VVPP] zu taggen.

Des Weiteren erscheint im ahd. Korpus öfter ein Partizip Präsens, im Einzelfall auch ein Partizip Präteritum, ohne finites Verb. Obwohl wir davon ausgehen können, dass dieses Partizip (wohl nach dem Vorbild des Lateinischen) eine Verbalhandlung repräsentiert und die VP bildet, kennzeichnen wir es um der Einheitlichkeit willen als Adjektiv [ADJD]. Dieses Vorgehen ermöglicht eine vergleichbare Suchabfrage aller Partizipien in nicht-attributiver Stellung.

In attributiver Stellung werden beide Partizipien selbstverständlich als [ADJA] gekennzeichnet.

Verb / Nomen

Der Infinitiv Präsens kann im ahd. Korpus nominalisiert und mit substantivischer Flexion versehen auftreten. Oftmals handelt es sich dabei um die Entsprechung des nhd. ‚*zu* – Infinitivs'. Wir orientieren uns an dem Vorbild des Nhd. und kennzeichnen die flektierten Infinitive immer als [VVINF], auch wenn ihre Funktion nicht in allen Fällen mit der des nhd. *zu* – Infinitivs übereinstimmt. Entsprechend wird die zum Infinitiv tretende Präposition *te* ‚zu' als [PTKZU] getaggt.

2.1.8 Kennzeichnung morphologischer Merkmale einer Wortform

Wie in 2.1.7 erwähnt, können in historischen Sprachstufen des Deutschen Wortarten flektieren, bei denen im nhd. Standard eine Flexion ausgeschlossen ist. Diese Formen werden selbstverständlich morphologisch annotiert, sofern eine overte Merkmalszuweisung vorliegt. Mit anderen Worten: es werden keine sogenannten Nullmorpheme gekennzeichnet, da nicht entscheidbar ist, ob es sich im Einzelfall tatsächlich um Nullmorpheme handelt, oder ob die unflektierte Grundform verwendet wird. Die Vergabe von morphologischen Kongruenzmerkmalen ist nicht immer einheitlich, ein prädikativ verwendetes Adjektiv z.B. kann flektieren oder nicht.

Prädikative Adjektive [ADJD] werden, sofern sie auf Partizipien zurückzuführen sind, ohne Steigerungsgrad getaggt, wodurch verbale Fügungen, die aus einer Kopula oder Auxiliar und einem Partizip bestehen und möglicherweise als Periphrase aufzufassen sind, systematisch gesucht werden können.

Folgende grammatische Kategorien können im Gegensatz zum Nhd. Flexionsmerkmale erhalten:

2.1.9 Änderungen an dem Morphologie-Annotationsschema hinsichtlich des Althochdeutschen

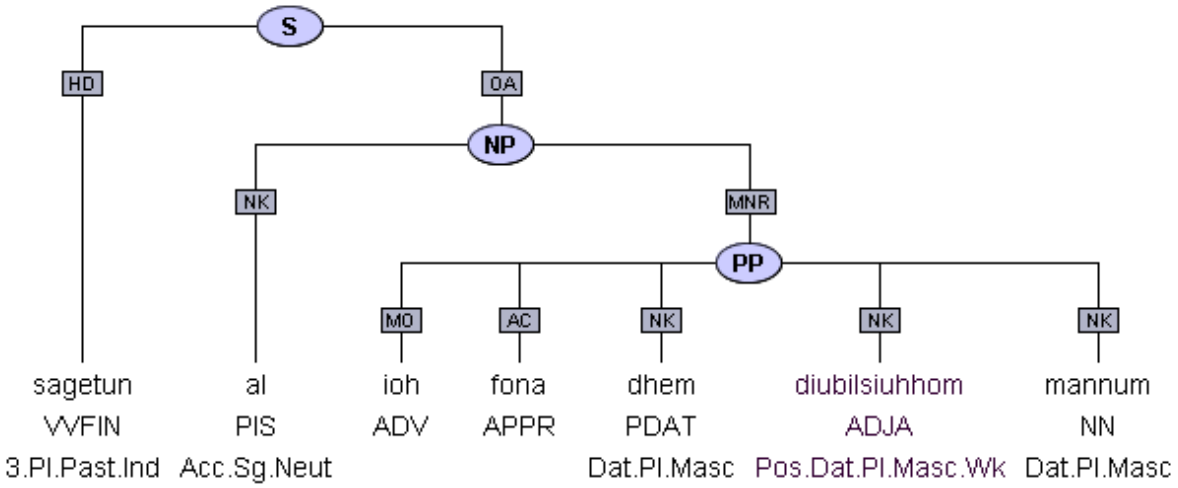
Aufgrund der vom Neuhochdeutschen etwas abweichenden Morphologie des historischen Deutsch war es notwendig, das Tagset zu ergänzen. Im Einzelnen handelt es sich dabei um folgende Tags:

<u>Flexionsstärke bei Adjektiven</u>

Die morphologischen Tags für die Adjektivflexion werden um einen Wert für die Flexionsstärke erweitert:

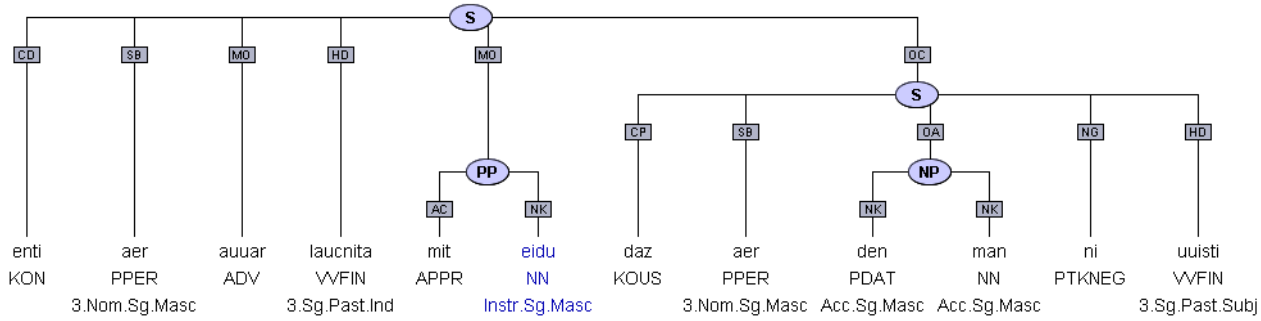
starke Flexion= (...)St

schwache Flexion= (...)Wk



Instrumentalis in der Nominalflexion

Als zusätzliches Kasusmerkmal wird bei Substantiven und Pronomina die Kategorie Instrumentalis mit dem Tag "Instr." vergeben.



2.2 Richtlinien für die Annotation des Mittelhochdeutschen

Die Annotation des Mittelhochdeutschen gliedert sich in folgende Ebenen:

- [Spec], Textwiedergabe nach der Edition
- [Word], Bearbeiteter Text
- [Bibl], Bibliographische Referenzebene
- [Lemma], Lemmatisierung der Wortformen
- [POS], Zuweisung von Wortart an die einzelne Wortform
- [Morph], Kennzeichnung morphologischer Merkmale einer Wortform
- [Syntax], Syntaktische Kommentierung

2.2.1 Text

Grundlage des mittelhochdeutschen Textkorpus ist das sogenannte *Speculum ecclesiae deutsch* (Staatsbibliothek München, Cgm 39), die älteste volkssprachliche Predigtsammlung des Deutschen.

Die Handschrift wird auf das Ende des 12. Jahrhunderts datiert und ist westbairischen Dialekts mit alemannischen Einsprengseln.

Die Predigten, die z.T. nach lateinischer Vorlage entstanden sind, behandeln verschiedene Feiertage des Kirchenjahres und geben so Vorlagen für den christlichen Gottesdienst. In der Regel wird eine Predigt durch ein lateinisches Bibelzitat eingeleitet, worauf sich die volkssprachliche Predigt anschließt.

Edition: Mellenbourn, Gert (1944). Speculum ecclisiae. Lunder Germanistische Forschungen 12. Lund: Gleerup.

2.2.2 Edition und Textwiedergabe

Die für das Korpus verwendete Edition erhebt den Anspruch, diplomatischen Grundsätzen zu folgen, kann diesen aber nach heutigen editorischen Vorgaben nicht gerecht werden.

Änderungen werden im fortlaufenden Text nach dem Vorbild lateinischer Vorlagen und abweichender Fragmente vorgenommen, wobei sämtliche Abweichungen von der edierten Handschrift durch Kursivierung gekennzeichnet werden, in Fußnoten wird die Schreibung der Handschrift verzeichnet.

Des Weiteren löst der Herausgeber die sehr häufigen Abkürzungen auf und kennzeichnet diese ebenfalls durch Kursivschreibung.

Diakritika werden in der Regel wie in der Handschrift wiedergegeben, der von der Hand γ verwendete Akzent und Besonderheiten von Hand α werden allerdings nicht dargestellt. Sich an Initialen anschließende Majuskeln werden ohne weitere Kennzeichnung durch Kleinschreibung ersetzt. Proklitische Partikel und Präpositionen werden getrennt, die Getrenntschreibung bei nominalen Komposita wird durch Zusammenschreibung ersetzt, beides wird ohne weiteren Verweis auf die Handschrift vorgenommen. Moderne Interpunktion samt Großschreibung am Satzanfang wird einfügt.

2.2.3 Bearbeitung

Die Edition wird zum der Zweck der weiteren Annotation wie folgt bearbeitet: Diakritische Diphthonge werden aufgelöst, wobei der diakritische Vokal als zweiter Bestandteil des Diphthongs angesehen wird, und das lange <ſ> wird durch ein rundes <s> ersetzt. Die Schreibung <v> für den Vokal /u/ wird durch <u> ersetzt. Kursivschreibung wird aufgelöst, ebenso wie die in der Edition erfolgte Zeilentrennung (auf der Editionsebene wiedergegeben durch <->).

2.2.4 Bibliographische Referenzebene

Wie im althochdeutschen Subkorpus beziehen sich die bibliographischen Angaben auf jeweils eine Zeile in der verwendeten Ausgabe. Sie bezeichnen 1. den Herausgeber der Ausgabe, 2. die Nummerierungsangabe der Predigt, 3. die Seitenzahl und 4. die Zeilenangabe, also z.B. *Mellbourn, 32/80,7*.

2.2.5 Lemmatisierung der Wortformen

Die Wortformen des mittelhochdeutschen Korpus werden anhand des Mittelhochdeutsch – Wörterbuchs von Matthias Lexer lemmatisiert. Dieses Wörterbuch führt die mittelhochdeutschen Lexeme in der (nach Lachmann) üblichen normalisierten Orthographie auf, so dass auf der Lemmaebene alle Stichworte problemlos abbfragbar sind.

2.3 Richtlinien für die Annotation des Frühneuhochdeutschen

Die Annotation des Frühneuhochdeutschen gliedert sich in folgende Ebenen:

- [Nuber], Textwiedergabe nach der Edition
- [Word], Bearbeiteter Text
- [Bibl], Bibliographische Referenzebene
- [Lemma], Lemmatisierung der Wortformen
- [POS], Zuweisung von Wortart an die einzelne Wortform
- [Morph], Kennzeichnung morphologischer Merkmale einer Wortform
- [Syntax], Syntaktische Kommentierung

2.3.1 Text

Das frühneuhochdeutsche Subkorpus beinhaltet einen Teil der 1544 gedruckten Abhandlung „Ein kurtze und einfeltige unterweisung (…)[11]“ des Regensburger Predigers Veit Nuber. Der Text ist bairisch und bietet in predigthafter Weise Erläuterungen zu biblischen Zitaten und Anweisungen zum Troste Kranker.

[11] Der vollständige Titel lautet: „Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen“. Eine Kopie des Drucks wurde uns freundlicherweise vom Projekt des Bonner Frühneuhochdeutsch-Korpus (http://www.korpora.org/Fnhd/) zur Verfügung gestellt.

Wir geben den Text weitestgehend originalgetreu wieder. Allerdings mussten wir aus technischen Gründen einige Änderungen vornehmen. So wird die Ligatur aus <t> und <z> aufgelöst, die Umlautkennzeichnung durch hochgestelltes <e> wird durch ein dem betreffenden Vokal nachgestelltes <e> ersetzt, ein mit einem Hochstrich versehenes <n> wird als <ñ> wiedergegeben und das häufig verwendete Druckerzeichen für ‚etcetera' wird als <etc.> dargestellt.

2.3.2 Bearbeitung

Den frühneuhochdeutschen Originaltext haben wir in verschiedener Hinsicht bearbeitet. Das lange <ſ> wird durch ein rundes <s> ersetzt. <dz> wird ergänzt als <daz>. Die Abkürzung <s.> vor Heiligennamen wird ausgeschrieben als <sankt>. Zusammengehörende Wortteile, die im Druck durch Zeilentrennung getrennt sind (im Druck und entsprechend auf der Textebene mit <-> dargestellt), werden als eine Wortform dargestellt. Das gilt auch für wenige Wortformen, deren Teile im Text durch Spatium getrennt wiedergegeben sind.

Zu diesen gehören im Einzelnen:

die weil	→	dieweil
zu kunft	→	zukunft
trost spruchen	→	trostspruchen
aus gericht	→	ausgericht

In einem Fall wird ein fehlendes Graphem ersetzt.

2.3.3 Lemmatisierung

Da bisher noch kein Gesamtwörterbuch speziell zum Frühneuhochdeutschen vorliegt, nutzen wir zur Lemmatisierung das *Deutsche Wörterbuch* von Jacob und Wilhelm Grimm.

3 Ressourcen- und Literaturverweise

Deutsche Diachrone Baumbank

http://korpling.german.hu-berlin.de/ddd/search.html

Annotationsschemata

Albert, S., J. Anderssen, R. Bader, S. Becker, T. Bracht, S. Brants, T. Brants, V. Demberg, S. Dipper, P. Eisenberg, S. Hansen, H. Hirschmann, J. Janitzek, C. Kirstein, R. Langner, L. Michelbacher, O. Plaehn, C. Preis, M. Pussel, M. Rower, B. Schrader, A. Schwartz, G. Smith and H. Uszkoreit (2005). TIGER-Annotationsschema. Tech. Rep., Universität Potsdam, Universität Saarbrücken, Universität Stuttgart.

(http://www.linguistics.ruhr-uni-bochum.de/~dipper/papers/tiger_annot.pdf)

Crysmann, Berthold; Hansen-Schirra, Silvia; Smith, George; Ziegler-Eisele, Dorothea (2005). TIGER Morphologie-Annotationsschema. Tech. Rep., Universität Potsdam, Universität Saarbrücken.

(https://files.ifi.uzh.ch/cl/siclemat/lehre/papers/tiger-morph.pdf)

Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999) Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical Report. Institut für maschinelle Sprachverarbeitung, Stuttgart.

(http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf)

Editionen

Hench, George Allison (1890). The Monsee Fragments. Strassburg: Karl J. Trübner

Mellenbourn, Gert (1944). Speculum ecclisiae. Lunder Germanistische Forschungen 12. Lund: Gleerup.

Nuber, Veit (1544). Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen. Regensburg.

Grammatiken

Braune, Wilhelm (2004[15]). Althochdeutsche Grammatik I. Bearb. Ingo Reiffenstein. Tübingen: Niemeyer.

Mettke, Heinz (1993[7]). Mittelhochdeutsche Grammatik. Tübingen: Niemeyer.

Paul, Hermann (2007[25]). Mittelhochdeutsche Grammatik. Bearb. Thomas Klein, Hans-Joachim Solms und Klaus-Peter Wegera. Tübingen: Niemeyer.

Schrodt, Richard (2004). Althochdeutsche Grammatik II. Tübingen: Niemeyer.

Wörterbücher

Grimm, Jacob und Wilhelm (2007). Deutsches Wörterbuch. Elektronische Ressource: http://germazope.uni-trier.de/Projects/WBB/woerterbuecher/dwb/wbgui?lemid=GA00001 Bereitgestellt durch: Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier.

Lexer, Matthias (2007). Mittelhochdeutsches Handwörterbuch. Elektronische Ressource: http://germazope.uni-trier.de/Projects/WBB/woerterbuecher/woerterbuecher/lexer/wbgui

Bereitgestellt durch: Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier.

Schützeichel, Rudolf (1995). Althochdeutsches Wörterbuch. Tübingen: Niemeyer.