XML-Kodierung des Bonner Frühneuhochdeutschkorpus

IKP-Arbeitsbericht NF 02

M. Diel, B. Fisseni, W. Lenders, H.-C. Schmitz

{mdi,bfi,wle,hcs}@ikp.uni-bonn.de

1 Das Frühneuhochdeutsch-Korpus (Auswahlkorpus)

Es gibt zwei Bonner Korpora des Frühneuhochdeutschen: 1. Ein ‚großes' Korpus bestehend aus 1500 Texten, hier das *Gesamtkorpus* genannt. 2. Eine Auswahl von 40 Texten aus dem Gesamtkorpus, verfügbar in maschinenlesbarer Form, hier das *Auswahlkorpus* genannt.

Das Gesamtkorpus wurde zwischen 1972 und 1974 an der Forschungsstelle *Frühneuhochdeutsch* der Universität Bonn zusammengestellt, um „in großer Breite die Textüberlieferung des 14.-17. Jahrhunderts für sprachgeschichtliche Untersuchungen verschiedenster Art bereit[zu]stellen" ([Hoffmann/ Wetter, 1985] S. XIV). Die Texte des Gesamtkorpus entstammen 22 Sprachlandschaften und 7 Zeitschnitten à 50 Jahre (1350-1700).

Das Auswahlkorpus entstand zwischen 1972 und 1985 im Rahmen des Projekts *Flexionsmorphologie des Frühneuhochdeutschen* unter Leitung von Werner Besch, Winfried Lenders (ab 1976), Hugo Moser und Hugo Stopp (bis 1981). Es diente als Materialgrundlage zur Analyse der Morphologie des Frühneuhochdeutschen und zur Erarbeitung mehrerer Bänden der *Grammatik des Frühneuhochdeutschen*, nämlich [Dammers et al, 1988], [Solms/ Wegera, 1991] und [Wegera, 1987].

Das Auswahlkorpus besteht aus zwei Teilen zu je 20 Texten. Die insgesamt 40 Texte entstammen 10 verschiedenen Sprachlandschaften (Tabelle 1). Jede Sprachlandschaft ist im Korpus durch vier Texte aus vier Zeitschnitten repräsentiert (Tabelle 2). Jeder Text hat eine dreistellige Nummer, die sich aus den Ziffern für das Teilkorpus, die Sprachlandschaft und den Zeitschnitt zusammensetzt. Gotthard Heideggers *Mythoscopia* beispielsweise gehört zum zweiten Teilkorpus, entstammt dem Osthochalemannischen, wurde 1698 verfasst, hat also die Nummer *217*. Die Texte sind nicht immer vollständig in das Korpus aufgenommen, sondern jeweils zu einem Ausschnitt von ca. 30 Normalseiten mit etwa 400 Wörtern. Der Umfang reicht aus, um die morphologische Repräsentativität der Texte zu

Teilkorpus	Sprachlandschaft (Nr)	Sprachlandschaft
1	1	Mittelbairisch (Wien)
1	2	Schwäbisch
1	3	Ostfränkisch (Nürnberg)
1	4	Obersächsisch
1	5	Ripuarisch (Köln)
2	1	Osthochalemannisch
2	2	Ostschwäbisch (Augsburg)
2	3	Elsässisch (Straßburg)
2	4	Hessisch
2	5	Thüringisch

Tabelle 1: Sprachlandschaften

Nr	Jahre
1	1350-1400
3	1450-1500
5	1550-1600
7	1650-1700

Tabelle 2: Zeitschnitte

gewährleisten. Drei Texte bestehen aus zwei Teilen (vgl. Tabelle 3). In den verfügbaren Versionen des Korpus ist von Text 117 nur der erste Teil – *Deo Gratias* – vertreten. Die Daten des zweiten Teils – *Mercks wol Soldat! Das ist: Die Glori von dem Heiligen Ritter Georgio [...]* – sind verloren. Genaueres zur Textauswahl findet sich in [Graser/ Wegera, 1978].

Die Texte des Auswahlkorpus sind durchgängig mit Wortklassen- und Formenbestimmungen (morphologischen Angaben) und sonstigen Angaben versehen. Die Annotierung ist im folgenden Kapitel 2 beschrieben.

117	Deo Gratias, Georgio
123	Eunuchus, Kommentar
127	Bericht, Chronik

Tabelle 3: Mehrteilige Texte

Das Auswahlkorpus steht in drei verschiedenen Versionen zur freien Verfügung: in der Original-kodierten Version, der XML-kodierten Version und der HTML-kodierten Version. Die vollständige XML-DTD und Beispiele für alle Kodierungen finden sich in Anhang A und B. Alle Versionen können von folgender Webseite bezogen werden:

`http://www.ikp.uni-bonn.de/dt/forsch/fnhd/`

Wer das Auswahlkorpus zitiert verweise bitte mit Angabe des URL: *Das Bonner Frühneuhochdeutsch-Korpus, IKP – Universität Bonn, http://www.ikp.uni-bonn.de/dt/forsch/fnhd/.*

2 Kodierungen

Die erste Version des (Auswahl-)Korpus wurde auf Lochkarten gespeichert, dann in den Zeichensatz der MS-DOS Codepage 437 übertragen. Die dadurch entstandene zweite Version nennen wir die *Original-kodierte Version*. Sie stimmt nicht in allen Punkten mit der von [Berg, 1982] beschriebenen ersten Version überein. Die Original-kodierte Version wurde vollständig in eine *XML-kodierte Version* transformiert. Maßgabe der Transformation war, die Kodierung in den neuen Standard zu überführen und dabei inhaltlich wenigst möglich zu verändern. Offensichtliche Fehler wurden korrigiert. Endlich wurde auf Grundlage der XML-kodierten Version eine ‚lesbare', an Information reduzierte, *HTML-kodierte Version* erstellt.

2.1 XML-kodierte Version

2.1.1 Annotation

Jede XML-kodierte Datei enthält eine Quellenangabe, wie sie zusätzlich im Quellenverzeichnis (s. Kapitel 3) kodiert ist, und den jeweiligen Quelltext. Quelltexte sind mit morphologischen und sonstigen Angaben versehen. Die Annotation ist in Anhang A definiert.

Texte sind in Seiten oder Blätter verschiedener Lagen (Foliozählung) gegliedert. Ein Text – 243: *Cube, Hortus Sanitatis* – ist in Kapitel gegliedert. Bei diesem ersetzt diese kapitelweise Gliederung die Aufteilung in Seiten oder Blätter. Seiten, Blätter und Kapitel sind als Elemente der Typen `seite`, `blatt` oder `kapitel` markiert. Die Elemente haben mehrere Attribute. Das Attribut `nr` hat als Wert die jeweilige Seiten-, Blatt oder Kapitelnummer. Das Attribut `lage` hat als Wert die jeweilige Nummer der Lage bei Foliozählung. Sofern Spalten vorhanden sind, hat `spalte` als Wert

die jeweilige Spaltennummer. Lagen und Spalten werden mit lateinischen Großbuchstaben nummeriert. Gemäß Konvention der Original-Kodierung beginnt in Text 137 mit jedem neuen Abschnitt eine neue Spalte, obwohl in der Originalschrift alle Abschnitte in der ersten Spalte beginnen (vgl. [Berg, 1982], S.48). Das Attribut `position` hat bei Foliozählung einen der Werte `recto` oder `verso`, standardmäßig den Wert `irrelevant`. Wert des Attributs `teil` ist bei mehrteiligen Quellen (vgl. Tabelle 3) der jeweilige Textteil. Zudem kann als Teil ein Vorwort oder Titelblatt angegeben sein. Ist kein Wert für das Attribut definiert, dann handelt es sich um den Haupttext der Quelle. Dies ist der Standardfall.

Seiten, Blätter und Kapitel sind in Zeilen aufgeteilt. Zeilen sind als Elemente vom Typ `zeile` markiert und nummeriert. Um konkurrierende Strukturen und Überschneidungen von Elementen zu vermeiden, sind Seiten, Blätter, Kapitel und Zeilen als leere Elemente (`<seite/>`, `<blatt/>`, `<kapitel/>`, `<zeile/>`), als Seiten-, Blatt-, Kapitel- und Zeilen*wechsel*, markiert. Absätze sind nicht markiert, auch nicht ohne weiteres aus der Zeilenkodierung zu erschließen.

Textstellen, die in der Original-kodierten Version als Eingriffe respektive Bearbeitungen, Hervorhebungen respektive Auszeichnungen, Überschriften oder Zitate (ohne Referenz) markiert sind, sind auch in der XML-Version markiert und zwar als `eingriff`, `emph`, `ueberschrift` oder `zitat`.

Alle Wortformen sind mit einem Tag `wortform` umschlossen. Das Element `wortform` hat mehrere Attribute. Das Attribut `typ` hat bei Adjektiven, Substantiven und Verben als Wert die jeweilige Wortklasse (`adjektiv`, `substantiv`, `verb`). Ist eine Wortform ein Adjektiv, ein Substantiv oder ein Verb, konnte aber aus dem Kontext nicht entschieden werden, zu welcher der drei Wortklassen es gehört, dann hat sein Attribut `typ` den Wert `potentiell`. Ist eine Wortform keiner der drei oben genannten Wortklassen zuzuordnen, dann hat sein Attribut `typ` den Wert `unbekannt`. Zu mehreren Wortformen sind im Annotationsteil des Original-kodierten Korpus zwei Wortklassen angegeben. Dabei handelt es sich zumeist um substantivierte Infinitive, die sowohl als Substantive als auch als Infinitive annotiert sind. Als Wert von `typ` wurde stets die erste Angabe übernommen. Der `typ` eines substantivierten Infinitivs ist demnach `substantiv`. Zweite Angaben wurde als Wert eines Attributs, dessen Namen mit `zweit` beginnt gespeichert. Der `zweittyp` eines substantivierten Infinitivs ist `verb`. Zweitannotationen betreffen nicht nur die Wortklasse. Außer dem Attribut `zweittyp` gibt es noch Attribute `zweittempus`, `zweitmodus` etc.

Sofern die Form eines Adjektivs oder Substantivs bestimmt werden konnte, sind Kasus, Numerus und Genus angegeben. Die entsprechenden At-

tribute des Elements `wortform` sind `kasus`, `numerus`, `genus`, ggf. auch `zweitkasus` etc. Adverbien sind als Adjektive mit dem Attribut-Wert-Paar `adverbial="ja"` kodiert. Sofern die Form eines Verbs bestimmt werden konnte, sind Tempus, Modus, Person, Form (Infinitiv, Partizip, finite Form) angegeben. Für die entsprechenden Attribute und deren Werte stehen wie üblich ‚sprechende' Namen. Flexive, ausgenommen Komparationssuffixe, sind als Wert des Attributs `flexiv` kodiert. Komparationssuffixe sind Werte des Attributs `komparation`. Genannt sind ferner das Lemma, ggf. das Prä- oder Suffix in normalisierter Form und der Vokal der Stammsilbe als Werte der Attribute `lemma`, `praefix`, `suffix`, `vokal`. Wert des Attribus `zeichen` ist ein folgendes Satzzeichen. Dieses Attribut ist nur unregelmäßig kodiert. Bei Fremdwörtern ist das Attribut `fremdwort` mit dem Wert `ja` belegt, `gefunden` ist die Wortform der Original-kodierten Version.

(Eigen-)Namen sind als solche markiert. Namen sind nicht mit Formbestimmungen versehen.

Zeichenketten, die ,,durch eine Leerstelle getrennt sind, bei der Bearbeitung aber als Einheit behandelt werden sollen" ([Berg, 1982] S. 40) sind in der Original-kodierten Version durch ein Doppelkreuz ,,#" verbunden. In der XML-kodierten Version ersetzt die Markierung `<trenn\>` das Doppelkreuz. Trennbare Verben und getrennte Substantive wie ,,Lust- und Kräutergärtchen" sind in der Original-kodierten Version durch einen Unterstrich an der Trennstelle (nur bei Substantiven) und eine ,,7" am Ende des zu vervollständigenden Wortes markiert: ,,Lust_ und Kräutergärtchen7" ([Berg, 1982] S.40). Die Kodierung ist in der XML-Version nicht umgesetzt. Zur Lemmatisierung wurden jeweils die vollständigen Formen verwendet: ,,Lustgärtchen" und ,,Kräutergärtchen". Das Lemma ist in beiden Fällen ,,Garten".

2.1.2 Zeichenkodierung

Die Zeichen der XML-Version sind gemäß dem Unicode™-Standard in UTF-8 kodiert. Auf Sonderzeichen wird mit Referenzen verwiesen. Referenzen haben ‚sprechende' Namen.

Alle Zeichen mit Diakritika sind aufgespalten. Sie bestehen aus Grundzeichen und darauf folgendem Diakritikum. Ein ,,$\ddot{a}$" ist somit nicht als einzelnes Zeichen kodiert (wie gemäß ISO-8859-1), sondern als ,,a" gefolgt vom Diakritikum ,,$\ddot{}$". Ein ,,$\overset{e}{a}$" ist kodiert als ,,a" gefolgt vom Diakritikum ,,$\overset{e}{\cdot}$". Nicht zu klären war die Bedeutung von übergeschriebenem ,,@" in der Original-kodierten Version. Es ist als übergeschriebenes ,,¬" wiedergegeben. In der Original-kodierten Version als nicht mit Sicherheit inter-

Textstellen	XML-Markierung	HTML-Markierung
Eingriff	`<eingriff>`	`<u>`
Hervorhebung	`<emph>`	`<i>`
Überschrift	`<ueberschrift>`	`<big><b>`
Zitat	`<zitat>`	`<b>`

Tabelle 4: Hervorhebungen (ohne Endtag)

pretierbar ausgezeichnete Superskripte sind in der XML-kodierten Version als Breve-Zeichen kodiert („$\overset{\smile}{\cdot}$“). Die Gleichbehandlung aller Zeichen mit Diakritika erlaubt die globale Suche nach solchen Zeichen, z.B. nach allen umgelauteten Buchstaben.

Auf den Lochkarten hat man keinen Unterschied zwischen verschiedenen Diakritika, die keine Buchstaben sind, gemacht. Man hat bei einem einzelnen Diakritikum (Punkt: $\dot{a}$. Akzent: $\acute{a}, \hat{a}$. Nicht aber beim Nasalkompendium oder beim überschriebenen „e“: $\bar{a}$, $\overset{e}{a}$) eine „1“ vor den Grundbuchstaben gesetzt („$1a$“), bei einem doppelten Diakritikum (Umlautpunkte: $\ddot{a}$ etc.) eine „2“ („$2a$“). Bei der Übertragung nach MS-DOS Codepage 437 hat man sich entschieden, alle so markierten einfachen Diakritika durch einen Akzent wiederzugeben und alle doppelten durch Umlautpunkte. Daraus ergeben sich Bezeichnungen von Wurzelverben in den Lemmatisierungen etwa als „*stán*“ und „*gán*“ anstatt als „*stân*“ und „*gân*“ (vgl. [Berg, 1982], S.30-35).

2.2 HTML-kodierte Version

In der HTML-kodierten Version ist die Identität jedes Textes durch einen Kurztitel festgestellt. Seiten, Blätter, Kapitel, Zeilen, Spalten und Lagen sind nummeriert. Textstellen, die in der XML-kodierten Version als Eingriffe, Hervorhebungen, Überschriften oder Zitate markiert sind, sind durch Unterstreichung, Kursivsetzung, Fett- oder Großschrift hervorgehoben (vgl. Tabelle 4). Zugunsten der Lesbarkeit wurde auf sonstige Annotationen verzichtet.

Die Zeichen der HTML-kodierten Version sind gemäß ISO-8859-1 (Latin 1) kodiert. Einge Sonderzeichen müssen umschrieben werden. Zeichen, die mit einem Diakritikum überschrieben sind und als solche nicht im Latin-1-Zeichensatz existieren, werden im HTML-Text hoch nachgestellt. So wird „$\overset{e}{a}$“ wird wiedergegeben als „a^e“. Superskripte, deren Bedeutung schon in

der Original-kodierten Version unsicher ist und die in der XML-kodierten Version durch ein Breve-Zeichen kodiert sind, erscheinen als „[?]“. Was in der Original-Version als übergeschriebenes „@“ und in der XML-Version als überschriebenes „¬“ kodiert ist, ist in der HTML-Version als nach- und hochgestelltes „@“ kodiert.

3 Quellenverzeichnis

Auch das Quellenverzeichnis ist in Original-kodierter, XML-kodierter und HTML-kodierter Version verfügbar. Die XML-Version basiert auf der mit den Quellenverzeichnissen in [Wegera, 1987] und [Hoffmann/ Wetter, 1985] abgeglichenen Original-Version. Ein Beispiel für eine Quellenangabe in den verschiedenen Versionen ist in Anhang C gegeben.

Die XML-kodierten Daten sind aufgeteilt in Basis- und Zusatzdaten. Die Basisdaten umfassen Kurztitel, Titel, Autor, Quelle (wenn nicht Original) und Herausgeber, Erscheinungsort, -jahr und andere Editionsspezifika, (Sprach-)Landschaft, (Entstehungs-)Zeit sowie die Anzahl der aufgenommenen Seiten. Die Zusatzdaten umfassen editorische Anmerkungen, eventuelle Vorlagen, Biographien der Verfasser, Übersetzer, Schreiber und Drucker, die Textart sowie eventuelle editorische Eingriffe (z.B. Kürzungen). Für die Elemente wurden ‚sprechende‘ Namen gewählt.

Angaben zu Landschaftszugehörigkeit, Zeitschnitt und Textgattung sind gemäß der Konventionen der *Forschungsstelle Frühneuhochdeutsch* (vgl. [Hoffmann/ Wetter, 1985] S. XVII f und XXIII f) in Siglen aufgelöst. Die Zeitschnitte sind kodiert wie in Tabelle 2 angegeben, es wurden lediglich römische statt arabische Zahlen verwendet. Die Siglen für Sprachlandschaften und Textgattungen sind in den Tabellen 5 und 6 angegeben. Die XML-DTD für das Quellenverzeichnis ist in die DTD für die Korpustexte integriert. Die DTD findet sich in Anhang A.

In der HTML-Version des Quellenverzeichnisses wurden die Angaben zum *Zeitraum* weggelassen und nur die präziseren Angaben zum *Jahr* gemacht. Die Siglen für Sprachlandschaften und Textgattungen wurden durch volle Angaben ersetzt.

A DTD für das Fnhd-Korpus und das Quellenverzeichnis

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- DTD
  zur XML-Version des Bonner Fruehneuhochdeutschkorpus, einer
  Umkodierung des EDV-Korpus Fruehneuhochdeutsch, das an der Bonner
```

Sigle	Sprachlandschaft
Mbair	Mittelbairisch (Wien)
Schwaeb	Schwäbisch
Ofr	Ostfränkisch (Nürnberg)
Obs	Obersächsisch
Rip	Ripuarisch (Köln)
Ohchal	Osthochalemannisch
Oschwaeb	Ostschwäbisch (Augsburg)
Els	Elsässisch (Straßburg)
Hess	Hessisch
Thuer	Thüringisch

Tabelle 5: Siglen für Sprachlandschaften

Sigle	Textgattung
KT	Kirchlich-theologischer Fachtext
CB	Chronikalischer und Berichtstext
EB	Erbaulicher Text
BI	Bibeltext
UN	Unterhaltsamer Text („schöne Literatur“)
PR	Privattext (Tagebuch, Brief etc.)
SP	(Meta-)Sprachtext (Grammatik etc.)
RG	Rechts- und Geschäftstext (Urkunde, Akte etc.)

Tabelle 6: Siglen für Textgattungen

```
  Forschungsstelle Fruehneuhochdeutsch fuer die Baende der Grammatik des
  Fruehneuhochdeutschen (H. Moser, H. Stopp, W.  Besch, Hrsg.) erstellt
  wurde.

  Vgl. hierzu u.a.
  Lenders, Winfried und Klaus-Peter Wegera (Hrsg.): Maschinelle Auswer-
    tung sprachhistorischer Quellen. Sprache und Information 3. Niemeyer,
    Tuebingen, 1982.
  Moser, Hugo, Hugo Stopp und Werner Besch (Hrsg.): Grammatik des
    Fruehneuhochdeutschen. Beitraege zur Laut- und Formenlehre, Bd. II-VI.
    Carl Winter Universitaetsverlag, Heidelberg.

Berhard Fisseni, 03/2002
-->

<!ELEMENT quelle (bibliographie, text) >

<!-- Zur Kodierung des Quellenverzeichnisses (Bibliographie)
  s. Datei quellenverzeichnis.readme -->

<!ELEMENT bibliographie         (Angabe_nr, datei, Basis, Zusatz?)>
<!ELEMENT       Basis
  (Kurztitel, Titel, Quelle?, Hrsg?, Ersch?, Landschaft, Zeit, aufgenommen)>
<!ELEMENT       Zusatz
  (Edit_Anm?, Vorlage?, Verfasser?, Uebersetzer?, Schreiber?, Drucker?,
  Textart, Edit_Eingr?)>
<!ELEMENT       Angabe_nr       (#PCDATA)>
<!ELEMENT       datei           (#PCDATA)>
<!ELEMENT       Kurztitel       (Autor?, Text)>
<!ELEMENT       Autor           (#PCDATA)>
<!ELEMENT       Text            (#PCDATA)>
<!ELEMENT       Titel           (#PCDATA)>
<!ELEMENT       Hrsg            (#PCDATA)>
<!ELEMENT       Quelle          (#PCDATA)>
<!ELEMENT       Ersch
  (Art?, Eort?, Ejahr?, Reihe?, Band?, Druck?, Umfang?)>
<!ELEMENT       Art             (#PCDATA)>
<!ELEMENT       Eort            (#PCDATA)>
<!ELEMENT       Ejahr           (#PCDATA)>
<!ELEMENT       Reihe           (#PCDATA)>
<!ELEMENT       Band            (#PCDATA)>
<!ELEMENT       Druck           (#PCDATA)>
<!ELEMENT       Umfang          (#PCDATA)>
<!ELEMENT       Landschaft      (Sprachraum, Ort?)>
<!ELEMENT       Sprachraum      (#PCDATA)>
<!ELEMENT       Ort             (#PCDATA)>
<!ELEMENT       Zeit            (philnr, Jahr)>
<!ELEMENT       philnr          (#PCDATA)>
<!ELEMENT       Jahr            (#PCDATA)>
<!ELEMENT       Edit_Anm        (#PCDATA)>
<!ELEMENT       Vorlage         (#PCDATA)>
<!ELEMENT       Verfasser       (#PCDATA)>
<!ELEMENT       Uebersetzer     (#PCDATA)>
```

```
<!ELEMENT       Schreiber       (#PCDATA)>
<!ELEMENT       Drucker         (#PCDATA)>
<!ELEMENT       Textart         (#PCDATA)>
<!ELEMENT       aufgenommen     (#PCDATA)>
<!ELEMENT       Edit_Eingr      (#PCDATA | ul)*>
<!ELEMENT       ul              (li+)>
<!ELEMENT       li              (#PCDATA)>
<!ATTLIST bibliographie textnr CDATA #REQUIRED>

<!-- Kodierung der Texte -->

<!ELEMENT text
     (#PCDATA|bibliographie|name|seite|blatt|kapitel|zeile|wortform|eingriff|
      ueberschrift|emph|zitat|trenn)*>
<!ELEMENT zitat (#PCDATA|wortform|zeile|eingriff|seite|blatt|kapitel|emph|trenn)*>
<!ELEMENT ueberschrift (#PCDATA|wortform|zeile|eingriff|seite|blatt|kapitel|name|emph)*>
<!ELEMENT eingriff (#PCDATA|zeile|ueberschrift|seite|blatt|kapitel|wortform|emph|trenn)*>
<!ELEMENT emph (#PCDATA|zeile|ueberschrift|seite|blatt|kapitel|wortform|name|trenn)*>
<!ELEMENT name (#PCDATA|zeile|seite|blatt|kapitel|wortform|trenn)*>
<!ELEMENT zeile EMPTY>
<!ATTLIST zeile nr CDATA #REQUIRED>
<!ELEMENT seite EMPTY>
<!ATTLIST seite
    nr CDATA  #REQUIRED
    spalte   CDATA     ""
    teil     CDATA    ""
>
<!ELEMENT blatt EMPTY>
<!ATTLIST blatt
    nr CDATA  #REQUIRED
    lage     CDATA         ""
    position (recto|verso|irrelevant) "irrelevant"
    spalte   CDATA     ""
    teil     CDATA    ""
>
<!ELEMENT kapitel EMPTY>
<!ATTLIST kapitel
    nr CDATA  #REQUIRED
    teil     CDATA    ""
>
<!ELEMENT trenn EMPTY>
<!ELEMENT wortform (#PCDATA|trenn|zeile)*>

<!--
     ZU DEN ATTRIBUTEN DES <WORTFORM>-TAGS:

     klasse:   Unterklassen sind nur bei Verben annotiert.
     zeichen:  Dieses Attribut enthaelt Satzzeichen, das nach dem Wort im
               Korpus steht, sofern es im Originalkorpus annotiert war.
     zweit...: Manche Woerter sind zweifach annotiert, z.B.
               substantivierte Infinitive primaer als Substantive, aber
               auch als Verben; diese sekundaeren Annotationen stehen in
               Attributen, die mit 'zweit' beginnen
```

```
-->

<!ATTLIST wortform
    adverbial   (ja|nein|unbekannt) "unbekannt"
    istprefix   (ja|nein) "nein"
    istverbal   (ja|nein) "nein"
    fremdwort   (ja|nein) "nein"
    form        (infinitiv|partizip|finit|unbekannt|irrelevant) "irrelevant"
    lemma       CDATA #IMPLIED
    gefunden    CDATA #IMPLIED
    praefix     CDATA #IMPLIED
    suffix      CDATA #IMPLIED
    vokal       CDATA #IMPLIED
    zeichen     CDATA #IMPLIED
    flexiv      CDATA #IMPLIED
    klasse      CDATA #IMPLIED
    komparation CDATA #IMPLIED
    komparationsstufe (superlativ|komparativ|positiv|unbekannt|irrelevant) "irrelevant"
    person      (1|2|3|unbekannt|irrelevant) "irrelevant"
    numerus     (singular|plural|unbekannt|irrelevant) "irrelevant"
    kasus       (nominativ|genitiv|dativ|akkusativ|unbekannt|irrelevant) "irrelevant"
    modus       (imperativ|indikativ|konjunktiv|unbekannt|irrelevant) "irrelevant"
    genus       (maskulinum|femininum|neutrum|unbekannt|irrelevant) "irrelevant"
    tempus      (praesens|praeteritum|unbekannt|irrelevant) "irrelevant"
    typ         (substantiv|adjektiv|verb|unbekannt|potentiell) "unbekannt"
    zweitadverbial   (ja|nein|unbekannt) "unbekannt"
    zweitfremdwort   (ja|nein) "nein"
    zweitform        (infinitiv|partizip|finit|unbekannt|irrelevant) "irrelevant"
    zweitklasse      CDATA #IMPLIED
    zweitlemma       CDATA #IMPLIED
    zweitgefunden    CDATA #IMPLIED
    zweitpraefix     CDATA #IMPLIED
    zweitsuffix      CDATA #IMPLIED
    zweitvokal       CDATA #IMPLIED
    zweitzeichen     CDATA #IMPLIED
    zweitflexiv      CDATA #IMPLIED
    zweitkomparation CDATA #IMPLIED
    zweitkomparationsstufe (superlativ|komparativ|positiv|unbekannt|irrelevant) "irrelevant"
    zweitperson      (1|2|3|unbekannt|irrelevant) "irrelevant"
    zweitnumerus     (singular|plural|unbekannt|irrelevant) "irrelevant"
    zweitkasus       (nominativ|genitiv|dativ|akkusativ|unbekannt|irrelevant) "irrelevant"
    zweitmodus       (imperativ|indikativ|konjunktiv|unbekannt|irrelevant) "irrelevant"
    zweitgenus       (maskulinum|femininum|neutrum|unbekannt|irrelevant) "irrelevant"
    zweittempus      (praesens|praeteritum|unbekannt|irrelevant) "irrelevant"
    zweittyp         (substantiv|adjektiv|verb|unbekannt|potentiell) "unbekannt"
    >


<!-- Entitaeten -->

<!ENTITY sect "&#167;"> <!-- Paragraphenzeichen -->

<!-- Diakritische Unicode-Zeichen, die ueber das vorherige Zeichen treten
```

```
-->
  <!ENTITY UML "&#776;"> <!-- Umlaut -->
  <!ENTITY AKUT "&#769;"> <!-- Akut -->
  <!ENTITY PUNKT "&#775;"> <!-- einzelner Punkt -->
  <!ENTITY barsuper "&#773;"> <!-- horizontaler Strich -->
  <!ENTITY asuper "&#867;">
          <!-- In Unicode 3.2 voraussichtlich a superscriptum-->
  <!ENTITY esuper "&#868;">
          <!-- In Unicode 3.2 voraussichtlich e superscriptum-->
  <!ENTITY isuper "&#869;">
          <!-- In Unicode 3.2 voraussichtlich i superscriptum-->
  <!ENTITY osuper "&#870;">
          <!-- In Unicode 3.2 voraussichtlich o superscriptum-->
  <!ENTITY usuper "&#871;">
          <!-- In Unicode 3.2 voraussichtlich u superscriptum-->
  <!ENTITY csuper "&#872;">
          <!-- In Unicode 3.2 voraussichtlich c superscriptum-->
  <!ENTITY dsuper "&#873;">
          <!-- In Unicode 3.2 voraussichtlich d superscriptum-->
  <!ENTITY hsuper "&#874;">
          <!-- In Unicode 3.2 voraussichtlich h superscriptum-->
  <!ENTITY msuper "&#875;">
          <!-- In Unicode 3.2 voraussichtlich m superscriptum-->
  <!ENTITY rsuper "&#876;">
          <!-- In Unicode 3.2 voraussichtlich r superscriptum-->
  <!ENTITY tsuper "&#877;">
          <!-- In Unicode 3.2 voraussichtlich t superscriptum-->
  <!ENTITY nsuper "&#878;">
          <!-- In Unicode 3.2 voraussichtlich n superscriptum-->
  <!ENTITY vsuper "&#879;">
          <!-- In Unicode 3.2 voraussichtlich v superscriptum-->
  <!ENTITY ssuper "&#830;">
          <!-- uebergeschriebene vertikale Tilde, steht fuer
               s superscriptum -->
  <!ENTITY atsuper "&#794;">
          <!-- uebergeschriebenes NICHT-Zeichen steht fuer
               @ superscriptum -->
  <!ENTITY fragsuper "&#728;">
          <!-- uebergeschriebener Halbkreis (Breve) steht fuer
               unbekanntes Superskript -->
  <!ENTITY oelig "&#339;">
          <!-- Ligatur aus o und e -->
  <!ENTITY OElig "&#338;">
          <!-- Ligatur aus O und E -->

<!ENTITY cent  "&#162;">

<!-- Latin1-Zeichen als Entitaeten.  Im XML-Text wurden kombinierte
     Zeichen (Zeichen mit Diakritika) prinzipiell nicht verwandt, sondern
     durch e&AKUT; oder a&UML; etc. wiedergegeben.
-->

<!ENTITY Agrave "&#192;">
```

```
<!ENTITY Aacute "&#193;">
<!ENTITY Acirc "&#194;">
<!ENTITY Atilde "&#195;">
<!ENTITY Auml "&#196;">
<!ENTITY Aring "&#197;">
<!ENTITY AElig "&#198;">
          <!-- Ligatur aus A und E -->
<!ENTITY Ccedil "&#199;">
<!ENTITY Egrave "&#200;">
<!ENTITY Eacute "&#201;">
<!ENTITY Ecirc "&#202;">
<!ENTITY Euml "&#203;">
<!ENTITY Igrave "&#204;">
<!ENTITY Iacute "&#205;">
<!ENTITY Icirc "&#206;">
<!ENTITY Iuml "&#207;">
<!ENTITY ETH "&#208;">
<!ENTITY Ntilde "&#209;">
<!ENTITY Ograve "&#210;">
<!ENTITY Oacute "&#211;">
<!ENTITY Ocirc "&#212;">
<!ENTITY Otilde "&#213;">
<!ENTITY Ouml "&#214;">
<!ENTITY Oslash "&#216;">
<!ENTITY Ugrave "&#217;">
<!ENTITY Uacute "&#218;">
<!ENTITY Ucirc "&#219;">
<!ENTITY Uuml "&#220;">
<!ENTITY Yacute "&#221;">
<!ENTITY THORN "&#222;">
<!ENTITY szlig "&#223;">
<!ENTITY agrave "&#224;">
<!ENTITY aacute "&#225;">
<!ENTITY acirc "&#226;">
<!ENTITY atilde "&#227;">
<!ENTITY auml "&#228;">
<!ENTITY aring "&#229;">
<!ENTITY aelig "&#230;">
          <!-- Ligatur aus a und : -->
<!ENTITY ccedil "&#231;">
<!ENTITY egrave "&#232;">
<!ENTITY eacute "&#233;">
<!ENTITY ecirc "&#234;">
<!ENTITY euml "&#235;">
<!ENTITY igrave "&#236;">
<!ENTITY iacute "&#237;">
<!ENTITY icirc "&#238;">
<!ENTITY iuml "&#239;">
<!ENTITY eth "&#240;">
<!ENTITY ntilde "&#241;">
<!ENTITY ograve "&#242;">
<!ENTITY oacute "&#243;">
<!ENTITY ocirc "&#244;">
```

```
<!ENTITY otilde "&#245;">
<!ENTITY ouml "&#246;">
<!ENTITY oslash "&#248;">
<!ENTITY ugrave "&#249;">
<!ENTITY uacute "&#250;">
<!ENTITY ucirc "&#251;">
<!ENTITY uuml "&#252;">
<!ENTITY yacute "&#253;">
<!ENTITY thorn "&#254;">
<!ENTITY yuml "&#255;">

<!-- Entitaeten, auf die im Quellenverzeichnis (Bibliographie) referiert wird. -->

<!ENTITY        ahe             "a&#868;">
<!ENTITY        ohe             "o&#868;">
<!ENTITY        uhe             "u&#868;">
<!ENTITY        uho             "u&#870;">
<!ENTITY        aine            "&#x00E6;"> <!-- = &aelig-->
<!ENTITY        alang           "a&#773;">
<!ENTITY        elang           "e&#773;">
<!ENTITY        nlang           "n&#773;">
<!ENTITY        olang           "o&#773;">
<!ENTITY        grad            "&#x00BA;">
<!ENTITY        tot             "&#x2020;">
```

B Beispiele für Kodierungen von Korpustexten

B.1 Der Anfang von Durandus' Rationale in XML-Kodierung

```
<seite nr="1"/>
<zeile nr="01"/>
<eingriff>
<ueberschrift>
<wortform gefunden="VORREDE" genus="femininum" kasus="nominativ"
   lemma="rede" numerus="singular" praefix="vor=" typ="substantiv"
   vokal="e">VORREDE</wortform>
<zeile nr="02"/>
<wortform gefunden="ZUM">ZUM</wortform>
<zeile nr="03"/>
<wortform gefunden="RATIONALE">RATIONALE</wortform>
<wortform gefunden="DIVINORUM">DIVINORUM</wortform>
<wortform gefunden="OFFICIORUM">OFFICIORUM</wortform>
</ueberschrift>
</eingriff>
<zeile nr="04"/>
<name><wortform gefunden="Aristotiles">Aristotiles</wortform></name>
<wortform gefunden="der">der</wortform>
<wortform gefunden="schreybet" form="finit" klasse="stark_a&PUNKT;"
```

```
   lemma="schreiben" modus="indikativ" numerus="singular" person="3"
  tempus="praesens" typ="verb">schreybet</wortform>
<wortform gefunden="in">in</wortform>
<wortform gefunden="dem">dem</wortform>
<wortform gefunden="pueche" genus="neutrum" kasus="dativ"
   lemma="buch" numerus="singular" typ="substantiv" vokal="ue">
pueche</wortform>
<wortform gefunden="von">von</wortform>
<wortform gefunden="der">der</wortform>
<wortform gefunden="auzrichtung" genus="femininum" kasus="dativ"
   lemma="richtung" numerus="singular" praefix="aus=" suffix="ung"
   typ="substantiv" vokal="i">auzrichtung</wortform>
<wortform gefunden="der">der</wortform>
<zeile nr="05"/>
<wortform gefunden="gemaine" genus="femininum" kasus="genitiv"
   lemma="ge-meinde" numerus="singular" typ="substantiv" vokal="ai"
   zeichen=":">gemaine</wortform>:
<zitat>
<wortform gefunden="We">We</wortform>
<wortform gefunden="dem">dem</wortform>
<wortform gefunden="lannde" genus="neutrum" kasus="dativ"
   lemma="land" numerus="singular" typ="substantiv" vokal="a"
   zeichen=",">lannde</wortform>,
<wortform gefunden="dez">dez</wortform>
<wortform gefunden="chu&esuper;nig" genus="maskulinum" kasus="nominativ"
   lemma="ko&UML;nig" numerus="singular" typ="substantiv" vokal="u&isuper;">
chu&esuper;nig</wortform>
<wortform gefunden="ein">ein</wortform>
<wortform gefunden="chind" genus="neutrum" kasus="nominativ" lemma="kind"
   numerus="singular" typ="substantiv" vokal="i">chind</wortform>
<wortform gefunden="ist" typ="verb">ist</wortform>
<wortform gefunden="und">und</wortform>
<wortform gefunden="des">des</wortform>
<wortform gefunden="fu&esuper;rsten" genus="maskulinum" kasus="nominativ"
  lemma="fu&UML;rst" numerus="plural" typ="substantiv" vokal="u&esuper;">
fu&esuper;rsten</wortform>
<wortform gefunden="frue">frue</wortform>
```

B.2 Der Anfang von Durandus' Rationale in HTML-Kodierung

```
<table>
  <tr>
   <td></td>
 </tr>
 <tr>
  <td colspan="2"> <br>Seite 1</td>
 </tr>
 <tr>
  <td>01</td>
  <td><u><big><b>VORREDE</b></big></u> </td>
 </tr>
 <tr>
```

```
 <td>02</td>
 <td><u><big><b>ZUM</b></big></u> </td>
</tr>
<tr>
 <td>03</td>
 <td><u><big><b>RATIONALE DIVINORUM OFFICIORUM</b></big></u></td>
</tr>
<tr>
 <td>04</td>
 <td>Aristotiles der schreybet in dem pueche von der auzrichtung der</td>
</tr>
<tr>
 <td>05</td>
 <td>gemaine: <b>We dem lannde, dez chu<sup>e</sup>nig ein chind ist
     und des fu<sup>e</sup>rsten frue</b> </td>
</tr>
```

B.3 Der Anfang von Durandus' Rationale in Original-Kodierung

```
|T111DurandusRationale
|A0001X|Z010 +V +U VORREDE
      @sg_Vorrede @sp_vor= @sl_rede @sk_112 @sv_e
|Z020 +V +U ZUM
|Z030 +V +U RATIONALE DIVINORUM OFFICIORUM
|Z040 +N Aristotiles -N der schreybet$ in dem pueche+ von der auzrichtung+ der
      @sg_ausrichtung @sp_aus= @sl_richtung @ss_ung @sk_312 @sv_i
      @sg_pueche @sl_buch @sk_313 @sv_ue
      @vg_schreybet @vl_schreiben @vk_3112 @vs_11a
|Z050 gemaine+: +Z We dem lannde, dez chu<enig ein chind ist und des fu<ersten frue
      @sg_chind @sl_kind @sk_113 @sv_i
      @sg_chu<enig @sl_könig @sk_111 @sv_u<i
      @sg_fu<ersten @sl_fürst @sk_121 @sv_u<e
      @sg_gemaine @sl_ge-meinde @sk_412 @sv_ai @sz_:
      @sg_lannde @sl_land @sk_313 @sv_a @sz_,
      @vg_ist
```

C Beispiel für Kodierungen von Quellenangaben

C.1 Quellenangabe zu Durandus' Rationale in XML-Kodierung

```
<Angabe>
 <Angabe_nr>1</Angabe_nr>
 <datei>111.xml</datei>
 <Basis>
  <Kurztitel>
   <Autor>Wilhelm Durandus</Autor>
   <Text>Rationale, Wien 1384</Text>
  </Kurztitel>
  <Titel>Durandus' Rationale in sp&auml;tmittelhochdeutscher &Uuml;bersetzung.
    Das vierte Buch nach der HS. CVP 2765</Titel>
```

```
  <Hrsg>G.H. Buijssen</Hrsg>
  <Ersch>
   <Eort>Assen</Eort>
   <Ejahr>1966</Ejahr>
   <Reihe>STTHG</Reihe>
   <Band>172</Band>
   <Umfang>369 S.</Umfang>
  </Ersch>
  <Landschaft>
   <Sprachraum>Mbair</Sprachraum>
   <Ort>vermutl. Wien</Ort>
  </Landschaft>
  <Zeit><philnr>I</philnr> <Jahr>1384</Jahr></Zeit>
  <aufgenommen>S. 1-39</aufgenommen>
 </Basis>
 <Zusatz>
  <Uebersetzer>vermutl. Leopold Stainreuter, *um 1340 in Wien, 1368 Augustiner dort,
   1377-1385 Lesemeister in Wien, Kaplan Herzog Albrechts III., &tot;um 1400 in Wien
  </Uebersetzer>
  <Textart>KT</Textart>
 </Zusatz>
</Angabe>
```

C.2 Quellenangabe zu Durandus' Rationale in HTML-Kodierung

```
<p>
<b>[<a href="111.html">Wilhelm Durandus: Rationale, Wien 1384</a>]</b></u>
<p>
<b>Durandus' Rationale in sp&auml;tmittelhochdeutscher &Uuml;bersetzung.
Das vierte Buch nach der HS. CVP 2765</b><br>
G.H. Buijssen (Hrsg), Assen 1966 (STTHG, Bd. 172), Umfang: 369 S.<br>
Sprachraum: Mittelbairisch (Wien), Ort: vermutl. Wien, Zeit: 1384<br>
Textgattung: Kirchlich-Theologischer Fachtext<br>
aufgenommen: S. 1-39.
<p>
&Uuml;bersetzer: vermutl. Leopold Stainreuter, *um 1340 in Wien,
1368 Augustiner dort, 1377-1385 Lesemeister in Wien, Kaplan Herzog Albrechts
III., &dagger;um 1400 in Wien<br>
```

C.3 Quellenangabe zu Durandus' Rationale in Original-Kodierung

```
1
DURANDUS, WILHELM: RATIONALE, WIEN 1384                        512111
1378  0349  ENR.: 111 ] KT ] HS ] Z2 ] 280 NS. ]
SEM. K 7452 / KOPIE
LEOPOLD STAINREUTER DURANDUS RATIONALE, WILHELM DURANDUS,
G.H. BUIJSSEN
TITEL: DURANDUS' RATIONALE IN SP2ATMITTELHOCHDEUTSCHER 2UBERSETZUNG.
DAS VIERTE BUCH NACH DER HS. CVP 2765 ( HRSG. V. ) G.H. BUIJSSEN.
ASSEN 1966 ( STTHG ) ( 172), 369 S.
HS., A
LANDSCHAFT: MBAIR., WOHL WIEN
```

```
ZEIT: 1, 1384
UEBERSETZER: WOHL LEOPOLD &STAINREUTER, *UM 1340 IN WIEN, 1368
AUGUSTINER DORT, 1377-1385 LESEMEISTER IN WIEN, KAPLAN HERZOG
ALBRECHTS III., +UM 1400 IN WIEN.
TEXTART: KT
VORLAGE: WILHELM &DURANDUS, RATIONALE.
AUFGENOMMEN: S. 1 - 39.
MBAIR/WIEN                  1/1384              1    KT   HS/ED
DER TEXT BEGINNT MIT DER NAECHSTEN ZEILE
```

C.4 Quellenangabe zu Durandus' Rationale nach [Hoffmann/ Wetter, 1985]

1291 [Leopold **Stainreuter**], Durandus' Rationale in spätmittelhochdeutscher Übersetzung. Das vierte Buch nach der Hs. CVP 2765. [Hrsg. v.] G.H. Buijssen. Assen 1966 (StThG), (172), 369 S.
ÜF: Hs., A LS: mbair., wohl Wien ZT: 1, 1384 VF: Übers. wohl Leopold Stainreuter, * um 1340 in W., 1368 Augustiner dort, 1377-1385 Lesemeister in W., Kaplan Herzog Albrechts III., † um 1400 in W. GA: KT VL: Wilhelm Durandus (1230/31-1296), Rationale

C.5 Quellenangabe zu Durandus' Rationale nach [Wegera, 1987]

Mbair I = G. H. Buijssen, Durandus' Rationale in spätmittelhochdeutscher Übersetzung. Das vierte Buch nach der Hs. CVP 2765, Assen 1966 (Studia Theodisca). [Wien 1384]
Ausgewählt: S. 1-39

Literatur

[Berg, 1982] Entwicklung eines Kodierungssystems am Beispiel frühneuhochdeutscher Texte. In [Lenders/ Wegera, 1982], S. 19-50. 2, 2.1.1, 2.1.2

[Dammers et al, 1988] Grammatik des Frühneuhochdeutschen. Beiträge zur Laut- und Formenlehre, Hg. von Hugo Moser, Hugo Stopp und Werber Besch, *Band IV: Flexion der starken und schwachen Verben*, Von Ulf Dammers, Walter Hoffmann und Hans-Joachim Solms, Heidelberg 1988. 1

[Graser/ Wegera, 1978] Graser, Helmut und Wegera, Klaus Peter (Hg.): Zur Erforschung der frühneuhochdeutschen Flexionsmorphologie, In *Zeitschrift für Deutsche Philologie* [ZfDPh] 97/1978, S. 74-91. 1

[Hoffmann/ Wetter, 1985] Hoffmann, Walter und Wetter, Friedrich (Bearb.): Bibliographie frühneuhochdeutscher Quellen. Ein kommentier-

tes Verzeichnis von Texten des 14.-17. Jahrhunderts (Bonner Korpus), Frankfurt am Main u.a. 1985. 1, 3, C.4

[Lenders/ Wegera, 1982] Lenders, Winfried und Wegera, Klaus Peter (Hg.): Maschinelle Auswertung sprachhistorischer Quellen, Sprache und Information 3, Tübingen 1982. 3

[Solms/ Wegera, 1991] Grammatik des Frühneuhochdeutschen. Beiträge zur Laut- und Formenlehre, Hg. von Hugo Moser, Hugo Stopp und Werber Besch, Band VI: Flexion der Adjektive, Von Hans-Joachim Solms und Klaus-Peter Wegera, Heidelberg 1991. 1

[Solms/ Wegera, 1998] Solms, Hans-Joachim und Wegera, Klaus-Peter: Das Bonner Frühneuhochdeutsch-Korpus. Rückblick und Perspektiven. In Bergmann, Rolf (Hg.), *Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum Ersten Göttinger Arbeitsgespräch zur Historischen Deutschen Wortforschung. 1. und 2. November 1996*, Stuttgart, Leipzig 1998, S. 22-39.

[Wegera, 1987] Grammatik des Frühneuhochdeutschen. Beiträge zur Laut- und Formenlehre, Hg. von Hugo Moser, Hugo Stopp und Werber Besch, Band III: Flexion der Substantive, Von Klaus-Peter Wegera. Heidelberg 1987. 1, 3, C.5