Bernhard Fisseni, Hans-Christian Schmitz & Bernhard Schröder

FndhC/HTML und FnhdC/S

Abstract deutsch

Wir beschreiben kurz den explorativen (Fnhd/HTML) und analytischen Zugang (Fnhd/S) zum Bonner Frühneuhochdeutsch-Korpus, auf das über http://www.ikp.uni-bonn.de/fnhd/ zugegriffen werden kann.[1]

Abstract englisch

We shortly describe an explorative (Fnhd/HTML) and an analytic interface (Fnhd/S) to the Bonn Corpus of Early New High German that can be accessed via http://www.ikp.uni-bonn.de/fnhd/.[2]

Es gibt zwei Bonner Korpora des Frühneuhochdeutschen, nämlich erstens ein aus 1500 Texten bestehendes, nicht digitalisiertes ‚großes' Korpus und zweitens ein aus 40 Texten bestehendes, digitalisiertes und annotiertes Teilkorpus. Das digitalisierte Teilkorpus entstand zwischen 1972 und 1985 im Rahmen des Projekts *Flexionsmorphologie des Frühneuhochdeutschen* und diente als Materialgrundlage für die Erarbeitung mehrerer Bände der *Grammatik des Frühneuhochdeutschen* [2]. Es enthält Ausschnitte von 40 Texten, die zusammen zehn Sprachlandschaften und vier Zeitschnitte repräsentieren – jeweils ein Textausschnitt pro Sprachlandschaft-Zeitschnitt-Paar. Die Textausschnitte umfassen je 30 Normalseiten mit rund 400 Wortformen.

Das digitale Teilkorpus enthält für einen Großteil seiner Wortformen morphologische Informationen: Für die Substantive, Verben und Adjektive sind u.a. die Lemmata, Wortbildung (Präfixe und Suffixe) und Wortklassen angegeben, bei nominalen Wortformen auch Kasus, Numeri und Genera, bei Verbformen Personen und Tempora, bei Adjektivformen Komparationsmorphologie und Komparationsstufen. Darüber hinaus sind Namen, Überschriften, Seiten- und Zeilenwechsel, Zitate, Eingriffe und Hervorhebungen als solche markiert.

Bis 2002 stand das digitalisierte Teilkorpus nicht zur allgemeinen Verfügung. Es lag nur lokal auf Disketten gespeichert vor. Außerdem folgte die Kodierung keinem allgemein anerkannten Standard und war weitgehend undokumentiert; dies erschwerte sowohl die Verarbeitung als auch die Interpretation der Daten oder machte sie gar unmöglich. Im Jah-

1 Das Korpus soll mittelfristig unter http://www.korpora.org/fnhd/ zugänglich sein.

2 In a medium-term perspective, the corpus will be accessible via http://www.korpora.org/fnhd/.

Sprache und Datenverarbeitung 1-2/2007: S. 67–69

re 2002 wurde die Kodierung nach XML transferiert, HTML-Versionen der Korpustexte wurden hergestellt, und die HTML-Texte wurden zusammen mit dem XML-kodierten und dem original kodierten Korpus im Internet veröffentlicht (vgl. [1]).

Die Veröffentlichung des Korpus wurde mit großem Interesse aufgenommen: Es gab häufige Zugriffe auf die entsprechende Webseite, außerdem gingen viele Anfragen bezüglich des Korpus bei uns ein. Allerdings war der Zugriff auf das Korpus für die Zielgruppe bei weitem nicht optimal gestaltet: Aus der HTML-Version waren die morphologischen Annotationen entfernt, so dass zwar die Texte gut, die linguistisch interessanten Zusatzinformationen aber nicht gelesen werden konnten. Das XML-kodierte Korpus konnte zwar durchsucht werden und war so einer linguistischen Analyse zugänglich, es stehen für XML wider Erwarten bis heute aber keine Werkzeuge für die Bewältigung dieser Aufgaben zur Verfügung, welche ohne Programmierkenntnisse zu bedienen sind. Daher ist offensichtlich, dass der bisherige Zugang für die linguistisch interessierten Nutzer wiederum inadäquat war.

Um die Verwendbarkeit des Korpus zu verbessern, haben wir nun zwei neue Zugänge zum Korpus geschaffen: Zum einen haben wir für die explorative Untersuchung HTML-Versionen aller Texte erzeugt (FnhdC/HTML), in denen die annotierte linguistische Information in leicht lesbarer Form enthalten ist. Legt man den Mauszeiger auf eine Wortform, so wird die vorhandene linguistische Information zu dieser Wortform in einem kleinen Fenster eingeblendet.[3] Zum anderen haben wir eine einfache Möglichkeit geschaffen, das Korpus zu durchsuchen (FnhdC/S). In einer Suchmaske können Muster für Wortformen oder Lemmata eingegeben werden, nach denen das Korpus durchsucht wird. Die Suche kann eingeschränkt werden: Zum einen kann die Suchbasis auf einzelne Texte, Sprachlandschaften oder Zeitschnitte sowie auf Textbereiche – Eingriffe, Hervorhebungen, Überschriften, Zitate – reduziert werden. Zum anderen kann die linguistische Annotation zur Eingrenzung der Suche verwendet werden. So können beispielweise gezielt alle Substantive im Genitiv Plural mit Genus neutrum gesucht und gefunden werden. Die Suchergebnisse werden in ihrem jeweiligen textuellen Kontext angezeigt. Die linguistische Annotation kann dabei sowohl für die Sucherergebnisse als auch für ihre Kontexte angezeigt werden. Die Funktionalität der Suchmaske wird auf der Suchmaske selbst erklärt; die Maske ist deshalb einfach zu benutzen.

Beim Entwurf der Schnittstelle FnhdC/S mussten wir einen Kompromiss zwischen ihrer Funktionalität und der Einfachheit ihrer Benutzung eingehen: Damit die Suchmaske einigermaßen übersichtlich ist, lassen wir nur eine Auswahl der möglichen linguistischen Merkmale zur Einschränkung von Suchanfragen zu. So ermöglichen wir beispielsweise nicht die Suche nach ausgewählten Präfixen oder Suffixen – eine Beschränkung, die unsere Erwartungen über die Nutzung des Korpus widerspiegelt. Sollte sich aber gerade eine solche Suchmöglichkeit als sinnvoll und wünschenswert herausstellen, kann die Suchmaske ohne weiteres angepasst werden. Darüber hinaus ist es derzeit nur möglich, nach einzelnen Wortformen, nicht aber nach Folgen von Wortformen zu suchen. Mögli-

3 Die Funktion verlangt, dass JavaScript im Browser aktiviert ist.

cherweise könnte das Fehlen einer syntaktischen Annotation (partiell) kompensiert werden, wenn man nach Folgen von Wortformen mit bestimmten Formbeschränkungen suchen könnte. Um eine solche Suche zu ermöglichen, müssten aber komplexe Suchmuster eingeführt werden. Die Bedienung der Suchmaschine würde dadurch wesentlich schwieriger und setzte nennenswerten Lernaufwand voraus. Es stellt sich die Frage, ob der potentielle Nutzer nicht besser gleich eine XML-taugliche, einfache Programmiersprache wie XSLT, Python oder gar Perl lernt und damit die seinen Interessen genügenden Funktionalitäten stets selbst erzeugen kann.

Es steht zu erwarten, dass die neuen Schnittstellen FnhdC/HTML und FnhdC/S zu verbessern und den Bedürfnissen ihrer Benutzer anzupassen sind – etwa indem komplexere Suchmuster zugelassen werden. Damit wir das tun können, sind wir auf Erfahrungsberichte und Anregungen angewiesen.

Literatur

[1] Diel, Marcel / Fisseni, Bernhard / Lenders, Winfried und Hans-Christian Schmitz (2002): XML-Kodierung des Frühneuhochdeutschkorpus, IKP-Arbeitsbericht NF 02, Universität Bonn. http://www.ikp.uni-bonn.de/ikpab/ikpab-nf02.pdf

[2] Moser, Hugo / Stopp, Hugo und Werner Besch (Hrsg.) (1988): Grammatik des Frühneuhochdeutschen, hrsg. Heidelberg 1970-1988.