DeutschDiachronDigital – Ein diachrones Korpus des Deutschen

Anke Lüdeling, Thorwald Poschenrieder und Lukas Faulstich

Abstract [engl.]: This paper describes plans for a diachronic corpus of German which contains texts from Old High German to Modern German. In order to serve as a resource for research questions from many different fields (linguistics, literature, lexicography etc.) the corpus must have a flexible architecture as well as a high degree of standardization of content. This flexibility is possible through a multi--layer standoff corpus model where the texts are stored in a central database. Standardisation is ensured through common tagsets on each annotation level.

Einführung

Der vorliegende Aufsatz hat die Konzeption eines diachronen Korpus des Deutschen zum Gegenstand. Dieses Korpus soll Texte von den althochdeutschen und altsächsischen Anfängen der Überlieferung bis – in sinnvoller Auswahl – hin zum älteren Neuhochdeutsch (bis um 1900) enthalten. Zu entwickeln ist es im entstehenden Projekt Deutsch-DiachronDigital (DDD), einer bundesweiten Initiative von Forscherinnen und Forschern aus der (historischen) Philologie, der (historischen) Sprachwissenschaft sowie aus Literaturwissenschaft, Korpuslinguistik und Informatik. [1]

[1] Dieser Beitrag beschreibt einen Planungszustand – kein fertiges Korpus. An der DDD-Initiative sind 12 Universitäten zuzüglich weiterer Forschungseinrichtungen aus Deutschland beteiligt; zusätzlich gibt es eine Reihe ausländischer Kooperationspartner. Der Förderungsantrag ist eingereicht, es gibt aber zum Zeitpunkt der Abgabe dieses Manuskripts (November 2004) noch keine Finanzierungszusage. Genaueres zum Projekt findet sich unter <http://www.deutschdiachrondigital.de>. Die beschriebenen Entscheidungen zu Korpuszusammensetzung, Annotationsebenen und dergleichen sind im Projekt im großen Kreise getroffen worden; das heißt, daß darin viele Fachleute eingebunden waren und wir hier Ergebnisse berichten, die nicht nur von uns getragen werden. Lediglich in Einzelfragen werden wir hier teilweise etwas genauer als

In diesem Beitrag beschreiben wir allgemein die Vision eines standardisierten diachronen Korpus des Deutschen, das für möglichst viele Nutzungsinteressen offenbleibt, und gehen dabei ganz speziell auf die bisher für das DDD-Projekt getroffenen Entscheidungen ein. Ausführlicher wird schließlich dargestellt, nach welchem Ablauf eine Quelle entsprechend den DDD-Maßgaben aufgearbeitet werden soll.

Motivation

Die Erstellung von historischen und diachronen Korpora ist sehr ressourcenintensiv und teuer. Das bedeutet, daß ein solches Korpus für möglichst viele unterschiedliche Interessen nutzbar sein muß.

Wir listen hier exemplarisch ein paar mögliche paläographische/typographische, lexikographische, sprachgeschichtliche, sprachwissenschaftliche, text- oder literaturwissenschaftliche Fragetypen, die an ein historisches und/oder diachrones Korpus herangetragen werden könnten, auf.

Ein Korpus als Textsammlung hat dabei gegenüber einer elektronisch vorliegenden Edition eines Textes den Zweck, daß die Texte miteinander vergleichbar sind. Das gilt sowohl innerhalb einer Sprachstufe als über Sprachstufen hinweg. Mit einem Korpus muß man qualitative genauso wie quantitative (statistische) Fragen beantworten können.

Synchrone qualitative Fragestellungen könnten zum Beispiel sein:

- Gehen Buchstabenalignierungen in Handschriften des Typs *xy* gemeinhin über Kompositionsfugen oder über die Fugen zwischen logischen Wortformen innerhalb zusammengeschriebener Textwortkomplexe hinweg?
- Wo weichen die Ausgänge kongruierender Adjektiv-Substantiv-Syntagmen in frühen althochdeutschen Texten ausdrucksseitig voneinander ab, wo hingegen sind sie homonym? Gibt es darunter heterographe mutmaßliche Homophone?

Synchrone qualitativ-quantitative Fragestellungen sind:

---

bereits abgestimmt. Das Korpus wird sukzessive aufgebaut; dabei stehen alle vorhandenen Texte jederzeit der Öffentlichkeit über einen Webbrowser zur Verfügung.

In einem Vorprojekt namens „Komplexe Datenbasen“ (finanziert von der Senatsverwaltung für Wissenschaft, Forschung und Kultur, Berlin) sind die Korpusarchitektur und die zugrundeliegende Datenbankstruktur entwickelt worden (siehe dazu Dipper et al. 2004).

- Welche Schreibvarianten und welchen Wortschatzreichtum (Type--Token-Verhältnis) bietet der Text *xy* im Vergleich zu seiner Parallelfassung *qz*?
- Welche Paradigmastellenbelegungen von Vollverb-Stämmen kommen in einem mittelniederdeutschen Text der Textsorte *xy* wie oft vor?
- In welchen strukturellen Merkmalen unterscheiden sich Textmengen aus Gedichten versus Prosatexten im Mittelhochdeutschen?
- In welchen Texten kommen Spielarten des Ortsnamens *xy* und des Personennamens *qz* zugleich vor?
- Lassen sich Wortschatzcharakteristika in der Romansprache mit weiblicher versus männlicher Urheberschaft festmachen?

Vorwiegend diachrone oder sprachvergleichende qualitative Fragestellungen könnten sein:

- Wie entwickelt sich die Schreibung der Wortformen zum *‚Ecke‘* bezeichnenden Hyperlemma über die hochdeutschen Sprachstufen hinweg?
- Welche Entsprechungen hat lat. *pulcher* in ahd. Übersetzungstexten, wenn die Adjektivform Personen kennzeichnet?

Diachrone qualitativ-quantitative Fragestellungen könnten sein:

- Wie stellt sich die Finitverb-Stellung in Hauptsätzen mit analytischen Prädikatformen altsächsisch versus altbairisch statistisch dar?
- Wie lang sind untergeordnete Teilsätze in altdeutschen autochthonen Texten gegenüber Übersetzungstexten?
- Nimmt die *variatio sermonis* vom Mittelhochdeutschen zum Frühneuhochdeutschen in geistlichen Texten zu?

Diese exemplarische Auswahl von Fragestellungen zeigt bereits, welche Eigenschaften ein diachrones Korpus haben muß: es muß vergleichbare Texte aus verschiedenen Sprachstufen/Genres etc. enthalten, man muß einzelne Teilkorpora auswählen können, und die Texte müssen mit Meta-Informationen annotiert sein.

Obwohl bereits viele historische Texte elektronisch vorliegen (für einen Überblick siehe Kroymann et al. 2004), können bisher kaum systematische Untersuchungen über verschiedene Texte einer Sprachstufe oder über Sprachstufen hinweg durchgeführt werden: die Digital-Texte sind nicht miteinander vergleichbar, da sie sich in Diplomatizität, Feinkörnigkeit der bibliographischen Angaben und anderen Annotationen unterscheiden; außerdem sind die verschiedenen Sprachstufen des Deutschen unterschiedlich gut abgedeckt. Eine

Hauptaufgabe von DDD ist es also, aus vielen unterschiedlichen Einzeltexten ein Korpus zu erstellen. Ein solches historisches Korpus muß einerseits eine möglichst flexible Architektur haben, damit jederzeit neue Texte und Annotationsebenen aufgenommen werden können und andererseits inhaltlich vieles standardisieren, damit vergleichende Untersuchungen möglich werden.

Beide Aspekte werden in den nächsten Abschnitten genauer dargestellt. Das Projekt ist stark interdisziplinär: in unserer Korpusarchitektur sind wir vor allem durch zwei Forschungsbereiche beeinflußt: die elektronische philologische Textverarbeitung (Computerphilologie, *Humanistic Text Processing*) und die Korpuslinguistik (Computerlinguistik, *Natural Language Processing*).[2]

Hintergrund: Computerphilologie[3]

In der Philologie sind in den letzten Jahrzehnten Methoden für die Volltextdigitalisierung (Retrodigitalisierung[4]) entwickelt worden, um Digital-Ausgaben zu erstellen. Gute Digital-Ausgaben sind keine reine Digitalisierung einer Textfassung, sondern verbinden sehr detailliert verschiedene Textfassungen miteinander und mit einem kritischen Apparat und manchmal mit weiteren Ressourcen, zum Beispiel mit Wörterbüchern (wie im Projekt „Mittelhochdeutsches Wörterbuch“[5]) oder mit einer Stemma-Berechnung (wie im Canterbury Tales Project; Blake/Robinson 1993–1997); in vielen Fällen sind Digital-Ausgaben seitenweise mit Digital-Faksimiles aligniert. Allgemein enthalten solche Digital-Editionen viel Wissen über einen einzelnen Text, aber wegen der fehlenden Standardisierung gibt es, wie bereits dargestellt, oft keine Vergleichsmöglichkeiten zwischen den Texten; diachrone Studien sind nur schwer möglich.

---

2 Diese Einteilung (in Anlehnung an Zampolli 2004) kann sicher nicht immer sauber durchgehalten werden.

3 Dieser Abschnitt ist kurz gehalten, weil wir davon ausgehen, daß die Methoden und Hintergründe der Leserschaft dieser Zeitschrift gut bekannt sind (für einen Überblick siehe Hockey 2001, Burch et al. 2003 oder die Beiträge zur Tagung *‚The State of the Art in Humanities Computing‘* im Jahresband 2003 dieser Zeitschrift). Wir konzentrieren uns daher mehr auf Methoden und Ziele die Korpuslinguistik.

4 Wir beschäftigen uns hier nicht mit einer reinen Faksimile-Digitalisierung.

5 Die URLs aller im Text erwähnten Korpusressourcen finden sich in den Referenzen.

In einem Korpus kann nicht jeder Text die gleiche sorgfältige philologische Behandlung erfahren, wie es in Einzel-Editionen möglich ist. Trotzdem übernehmen wir aus der Computerphilologie hohe Ansprüche an die Diplomatizität (Urkundentreue), ferner die Möglichkeiten, zu einem Text weitere Informationen hinzuzufügen, etwa Digital-Faksimiles zu alignieren.

Hintergrund: Korpuslinguistik

Während sich die philologische Textverarbeitung zumeist auf die sehr detaillierte Erfassung und Beschreibung einzelner Texte (oder Werke eines bestimmten Autors) bezieht, beschäftigt sich die Korpuslinguistik mit Textsammlungen. Dabei liegt der Schwerpunkt eindeutig auf der (automatischen) Verarbeitung großer Textmengen aus modernen Sprachstufen – auch wenn natürlich historische Korpora und andere sogenannte *special corpora* (im Sinne von Sinclair 1996; s.u.) bestehen.[6] Es gibt viele Definitionen von „Korpus", wir beziehen uns hier auf eine relativ eng gefaßte von Sinclair (1996):

> A corpus is a collection of pieces of language that are selected according to explicit linguistic criteria in order to be used as a sample of the language […] A computer corpus is a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origin and provenance.

Aus dem Zitat wird deutlich, wie sich Korpora von Digital-Ausgaben unterscheiden: Ein Korpus besteht aus mehreren Texten, die nach vorgegebenen Kriterien ausgewählt, und – im Gegensatz zur Digital--Ausgabe – standardisiert und mit weiteren Angaben versehen (annotiert) sind. Bevor wir im Folgenden die einzelnen Bereiche Korpuszusammensetzung, Annotation und Auswertung in DDD genauer besprechen, möchten wir kurz einige Eigenschaften von großen Korpora beschreiben und zeigen, welche wir übernehmen.

Bisher sind die meisten Korpora in einer flachen Datei gespeichert, in der alle Annotationen an einzelnen Wörtern (Tokens) hängen. Jede Annotationsebene (also etwa Lemma oder Wortart; siehe nächsten Abschnitt) ist durchgängig für das ganze Korpus annotiert. Bei großen Korpora moderner Sprachstufen erfolgt die Annotation automatisch,

---

[6] Das erste elektronisch vorliegende Korpus war ein historisches (Busa 1974–1980).

wobei eine gewisse Fehlerrate in Kauf genommen wird. Für DDD wäre eine solche flache Dateistruktur nicht passend: es muß möglich sein, jederzeit Annotationsebenen hinzuzufügen, ohne dabei die bestehenden Ebenen zu stören; und es muß möglich sein, ausgewählte Teile eines Korpus mit einer gewissen Annotationsebene zu versehen. In den letzten Jahren sind mit den sogenannten Stand-off-Korpora passendere Korpusmodelle entwickelt worden.[7] In Stand-off-Korpora werden die Daten in einer Referenzdatei (Timeline) gespeichert; die Annotationsebenen sind dann getrennte Dateien, die jeweils auf bestimmte Stellen in der Referenzdatei verweisen.

Korpuszusammensetzung

Die Zusammensetzung eines Korpus bestimmt, wofür das Korpus eingesetzt werden kann. Da das DDD-Korpus für viele unterschiedliche Forschungsfragen genutzt werden soll, muß die Zusammensetzung möglichst ‚repräsentativ' sein (der Begriff ist insofern problematisch, als man die Grundgesamtheit ja nicht bestimmen kann; für verschiedene Fragestellungen kann man unterschiedliche Repräsentativitätsbegriffe entwickeln, siehe zum Beispiel Klein 1991, Biber 1993, Solms/Wegera 1998). Außerdem muß die Möglichkeit offenbleiben, jederzeit weitere Texte hinzuzufügen.

Für vergleichende Untersuchungen – seien dies Untersuchungen zum Sprachwandel, Genrevergleiche oder auch lexikographische Untersuchungen – ist eine Vergleichbarkeit über verschiedene Sprachstufen hinweg nötig; im Idealfall sollte sich also nur ein einziger Parameter (zum Beispiel Zeit) unterscheiden, während alle anderen Parameter (wie Textsorte, Formalisierungsgrad) gleichbleiben. Das ist natürlich bei älteren Sprachstufen schwerer möglich als bei heutigen: zum einen, weil sich viele Textsorten (wie Roman oder Tageszeitungsberichte) erst später entwickelt haben und andere (wie Evangelienharmonien) verschwunden sind, und zum anderen, weil viel weniger Material erhalten ist. Man bekommt also immer nur in

---

[7] Stand-off-Modelle sind vor allem für multimodale Korpora entwickelt worden, in den ein Sprachsignal mit seiner Transkription und eventuell noch weiteren Informationen aus anderen Modi wie z.B. Gestik aligniert werden muß. Das Sprachsignal bildet den Zeitstrahl (Timeline), auf den sich alle weiteren Ebenen beziehen. Im NITE-Projekt, an das wir uns anlehnen, wird eine allgemeine Architektur für Stand-off-Korpora entwickelt (Carletta et al. 2003). Viele Stand-off-Korpora sind in XML kodiert.

Teilbereichen Kontinuität. Wenn man eine Matrix aller relevanten Parameter aufstellt (beispielsweise Sprachstufe, Textsorte und Dialekt), bleiben besonders bei historischen Korpora zwangsläufig einige Zellen leer (dazu siehe zum Beispiel die Diskussionen in Rissanen et al. 1993).

In DDD werden die Texte nach den Parametern Zeit, Dialekt und Textsorte ausgewählt. Die daraus entstehende Matrix ist allerdings nur für die Sprachstufen Mittelhochdeutsch und Frühneuhochdeutsch wirklich anwendbar. Davor gibt es zu wenige Texte, daher werden alle größeren althochdeutschen und altsächsischen Texte aufgenommen. Danach gibt es zu viele Texte, daher beschränken wir uns im älteren Neuhochdeutschen zunächst auf die drei hochsprachlichen Textsorten Roman, Zeitung und Brief.

In Sinclairs Definition oben ist von *pieces of language* die Rede, nicht von „Texten". In manchen Korpora werden bewußt statt ganzer Texte nur Textausschnitte einer gewissen Länge aufgenommen, damit diese direkt miteinander verglichen werden können. DDD hat dagegen die Entscheidung getroffen, wenn möglich, immer ganze Texte aufzunehmen. Bei der Abfrage können dann beliebig lange Textausschnitte ausgewählt werden.

Annotation

Für viele Fragestellungen ist es entscheidend, daß man die Sprachdaten annotiert, also mit Metadaten versieht.

In der Korpuslinguistik werden meist drei Typen von Annotationen unterschieden: Header-Annotationen, positionelle Annotationen und strukturelle Annotationen (siehe zum Beispiel Evert/Fitschen 2004) – in unserem Korpusmodell sind alle Annotationen zur Struktur auch positionell.[8]

Header-Annotationen sind Auszeichnungen zu einem ganzen Text im Korpus. Dazu gehören etwa grundsätzliche Angaben über Textgeschichte, über Urheberschaft und Textsorte, über die Schreiberhände oder das Vorkommen von Sonderzeichen genauso wie

[8] Die Unterscheidung von Annotationen in „strukturell" und „positionell" beruht auf einem flachen Korpusmodell, in dem die Tokens als sogenannte Textpositionen geführt wurden. Positionelle Annotationen hängen dann jeweils an einem oder mehreren Tokens, während strukturelle zwischen den Tokens stehen. Diese Terminologie paßt technisch nicht zu unserem Stand-off-Modell: wir behandeln alle Annotationen positionell (Strukturangaben beziehen sich hier also auch immer auf Zeichen oder Zeichenketten).

Angaben über Vorverarbeitungsstandards und -werkzeuge. Viele dieser Angaben werden sinnvollerweise strukturiert, der Rest als Klartext beigefügt. Es gibt hier bereits sehr detaillierte Standards von der Text Encoding Initiative, auch umgesetzt im Corpus Encoding Standard von EAGLES, an die sich das DDD Projekt halten wird. Die erlaubten Annotationswerte werden im Projekt unter Rückgriff auf etablierte oder sich entwickelnde Standards, zum Beispiel ISO/TC 37/SC 4 standardisiert.

Detaillierte Header-Informationen ermöglichen die Zusammenstellung von Subkorpora (zum Beispiel alle Texte aus dem 16. Jahrhundert oder alle Briefe, die von Frauen geschrieben wurden). Dies ist für viele vergleichende Untersuchungen notwendig.

Positionelle Annotationen sind Angaben zu einer bestimmten Korpusposition. Im Gegensatz zu den meisten bisher üblichen Korpora ist bei uns die Bezugsgröße nicht ein Token (also in etwa ein graphemisches Wort), sondern ein Zeichen. Für jeden Typ von Angaben wird eine Annotationsebene definiert. Für jede Annotationsebene gibt es ein Tagset, das die möglichen Werte spezifiziert, und Annotationsrichtlinien. Im Prinzip kann es in unserem Korpusmodell beliebig viele Annotationsebenen geben. Beispiele sind eine paläographische Ebene mit Angaben, die sich auf die Schriftzeichen beziehen, wie etwa Initialbuchstabe, Schriftfarbe usw.; eine Ebenen zur physischen Struktur, die Zeilen, Seiten etc. markiert; eine Ebene zur logischen Struktur, die Sätze, Absätze etc. angibt; eine Ebene zur Lemma-Annotation von Wortformen etc.

Eine Sonderform der positionellen Auszeichnung ist die sog. Alignierung, bei der die Annotation nicht aus der Zuweisung einer metatextlichen Angabe besteht, sondern aus der In-Bezug-Setzung einer oder mehrerer Textspannen der einen Textebene mit einer oder mehreren entsprechenden Textspannen einer anderen Textebene. Beispiele sind die Alignierung von lateinischem Original und althochdeutscher Entsprechung in Interlinear-Übersetzungen oder die Alignierung von Digital-Texten mit Digitalfaksimile-Abschnitten.

Wie erwähnt, können in DDD im Prinzip beliebig viele Annotationsebenen eingeführt werden. Die Architektur kann Texte mit unterschiedlichen Auszeichnungsebenen verarbeiten. Um vergleichende Untersuchungen zu ermöglichen, haben wir uns jedoch auf einige Standards geeinigt: die meisten Texte (das sogenannte Kernkorpus) werden mit Lemmanamen, Wortart und Flexionsmorphologie annotiert. Dabei werden wir uns für die Wortarten und Flexionsmorphologie möglichst an das Stuttgart-Tübingen-Tagset STTS anlehnen (Schiller et

al. 1995). Die Lemma-Annotationen sind problematischer: jede Sprachstufe wird eine eigene Normalisierung und Abbildung auf Lemmata vornehmen. Zusätzlich ist ein Hyperlemma-System vorgesehen, das die Lemmata der verschiedenen Sprachstufen miteinander in Beziehung setzt. Dies ist schwierig, da etymologische und semantische Beziehungen zwischen den Lemmata sich widersprechen können (siehe dazu Gévaudan 2002).

Im Gegensatz zur automatischen Annotation für Korpora moderner Sprachen werden wir im DDD-Projekt manuell oder semi-automatisch annotieren. Dies liegt zum einen an fehlenden Ressourcen wie elektronischen Lexika und an der großen Unterschiedlichkeit der Texte (dies erschwert sowohl regelbasierte als auch statistische Verfahren) und zum anderen an den hohen Qualitätsansprüchen an ein historisches Korpus: Fehlerraten von 5% oder mehr sind nicht akzeptabel.

Technische Architektur des Korpus

In diesem Absatz wird die technische Architektur kurz angerissen. Abbildung 1 zeigt die geplante Systemarchitektur. Wir sehen eine web-basierte Client-Server-Architektur vor, um die technischen Zugangsvoraussetzungen für die Benutzer und Benutzerinnen möglichst niedrig zu halten. Für die Abfrage des DDD-Korpus wird zunächst nur ein Web-Browser und eventuell ein PDF-Viewer benötigt. Bearbeiter von Texten des Korpus benötigen prinzipiell nur einen allgemeinen XML-Editor. Es soll jedoch ein für das verwendete XML-Format speziell angepaßter Editor im Rahmen des Projekts bereitgestellt werden, welcher eine komfortablere und besser geführte Eingabe dieses Formats erlaubt. Ebenso sollen für die grammatikalische Annotation übliche Annotationswerkzeuge genutzt und – falls erforderlich – geeignet angepaßt werden. Die offline bearbeiteten Texte werden mittels Web-Browser zum Server hochgeladen und durch geeignete Import-Module in die Korpus-Datenbank eingepflegt, die auf einem zentralen Server in einem relationalen Datenbanksystem vorgehalten wird. Das zugrundeliegende Datenmodell und die Konversion in andere Formate werden genauer in Dipper et al. (2004) beschrieben.

Auf diese Datenbank kann über einen Web-Server zugegriffen werden, wobei die Such-, Import- und Export-Funktionalität durch zwischengelagerte Module in der Anwendungslogik-Schicht implementiert wird.

Wir sehen unterschiedlich komplexe Suchoberflächen für unterschiedliche Nutzergruppen – von Gelegenheitsnutzern bis hin zu Expertennutzern – vor. Die Anforderungen an diese Suchfunktionalität werden weiter unten noch genauer besprochen. Technische Lösungsansätze dafür werden in Faulstich et al. (2004) diskutiert. Die Export-Module stellen die Texte des Korpus bzw. Ausschnitte davon inklusive wählbarer Annotationsschichten in unterschiedlichen Dokumentformaten bereit. Als primäre Formate sehen wir XHTML für die Bildschirmrepräsentation, PDF für Druck und Offline-Präsentation sowie ein TEI-konformes XML-Format für die Offline-Analyse und Bearbeitung vor. Zusätzliche Import- und Export-Formate können durch Hinzufügen entsprechender Module unterstützt werden.

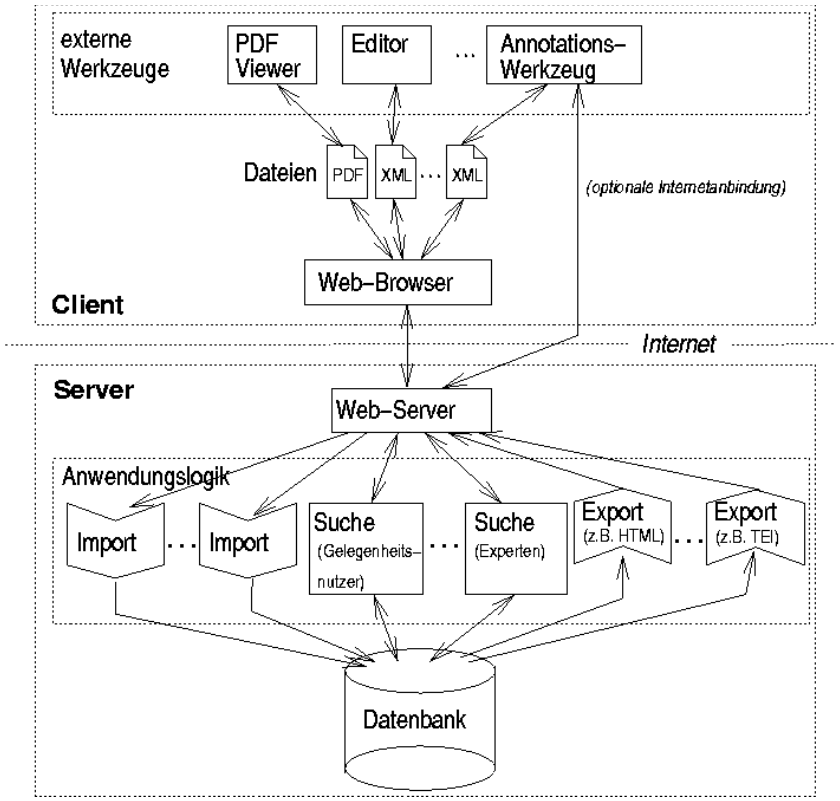

Abbildung 1: Systemarchitektur

Suche und Auswertung

Im Gegensatz zu Digital-Ausgaben, bei denen die Suche nach einzelnen Textstellen über eine Volltextsuche im Vordergrund steht, erfordern die umfangreichen Annotationen eines linguistischen Korpus wie des DDD--Korpus ungleich komplexere Suchmöglichkeiten. Im allgemeinen sucht man mit einer Anfrage nach Textabschnitten und Annotationen, die in einer bestimmten Beziehung zueinander stehen; sollen zum Beispiel alle Sätze gefunden werden, in denen das Wort mit dem Lemmanamen *„sagen"* in der 1. Person Plural auftritt, so setzt das Suchergebnis für jeden Treffer jeweils eine Satz-Annotation *s* , eine Wort-Annotation *w* , eine Lemma-Annotation *l* und eine flexionsmorphologische Annotation *f* miteinander in Beziehung, wobei folgende Bedingungen erfüllt sein müssen:

- Die Textspanne von *s* enthält die Textspanne von *w* , bzw. *w* ist eine Unterannotation von *s* .
- Die Lemma-Annotation *l* bezieht sich auf die Wort--Annotation *w* und hat den Wert *„sagen"*.
- Die flexions-morphologische Annotation *f* bezieht sich auf die Wort-Annotation *w* . Das Attribut *f.person* hat den Wert „1", und das Attribut *f.numerus* hat den Wert „Plural".

Damit ergeben sich folgende Anforderungen an eine angemessene Suchfunktionalität für DDD:

1. Eine Vorauswahl von Texten zu einem Teilkorpus muß über Bedingungen auf den Header-Angaben möglich sein.
2. Für vergleichende Untersuchungen ist oft ein ausgewogenes Korpus notwendig. Dazu sollen Texte für die Suche auf eine bestimmte einheitliche Länge eingeschränkt werden können.
3. Einzelne Zeichenketten und Reguläre Ausdrücke müssen auf allen Textebenen suchbar sein.
4. Annotationen müssen nach Typ und anhand auf ihre Attributwerte bezogener Bedingungen auswählbar sein.
5. Textspannen und Annotationen müssen miteinander in Beziehung gesetzt werden können:
    - einander enthaltende, direkt aufeinander folgende oder sich überschneidende Textspannen;
    - Textspannen in einer Textebene mit korrespondierenden Textspannen in einer alignierten Textebene;
    - eine Annotation mit der durch sie annotierten Textspanne;
    - eine Annotation mit ihren hierarchischen Nachkommen

(Unterannotationen) bzw. Vorfahren, beispielsweise ein Paragraph eines Texts mit den darin enthaltenen Sätzen und Wörtern, aber auch mit dem Kapitel, in dem er selbst enthalten ist.

Wie in Faulstich et al. 2004 näher beschrieben, orientieren wir uns beim Entwurf der Anfragesprache und Suchoberflächen an existierenden Lösungen zur Korpusanfrage wie am Corpus Query Processor (Christ 1994) und TigerSearch (Brants et al. 2002, Lezius 2002).

Zusätzlich zur Suche sollen auch quantitative Auswertungen (Stichwort: „deskriptive Statistik") unterstützt werden, zum Beispiel die relative Häufigkeit bestimmter Lemmata abhängig vom Texttyp oder die Erkennung statistisch auffälliger Muster wie zum Beispiel Kollokationen (vergl. Evert 2004) und von Trends, so etwa im Lauf der Zeit in Mode kommende Konstruktionen.

Bei der Suche nach Textspannen und Annotationen, aber auch bei quantitativen Untersuchungen müssen die Ergebnisse geeignet präsentiert werden. Suchergebnisse müssen im Kontext der zugrundeliegenden Texte (im Sinne von Konkordanzen) ausgegeben werden, wobei der Benutzer die engere Text-Umgebung eines Treffers nicht nur sehen, sondern von dort ausgehend auch den gesamten Text erkunden können soll. Es sind verschiedene Darstellungsarten denkbar, die vom Benutzer wähl- und anpaßbar sein müssen. Insbesondere müssen die unterschiedlichen Text- und Annotations- und Bildebenen parallel oder alternativ sichtbar gemacht werden können und miteinander geeignet verlinkt sein. Zur Anzeige quantitativer Ergebnisse müssen entsprechende Tabellen oder Graphiken generiert werden.

Arbeitsablauf bei der Aufarbeitung eines Textes in DDD

In diesem Abschnitt schildern wir grob die DDD-Bearbeitungsschritte, welche ein in das Korpus einzugliedernder Text nach jetziger Planung idealerweise erfahren soll; schematisch ist dies in Abbildung 2 zusammengefaßt.

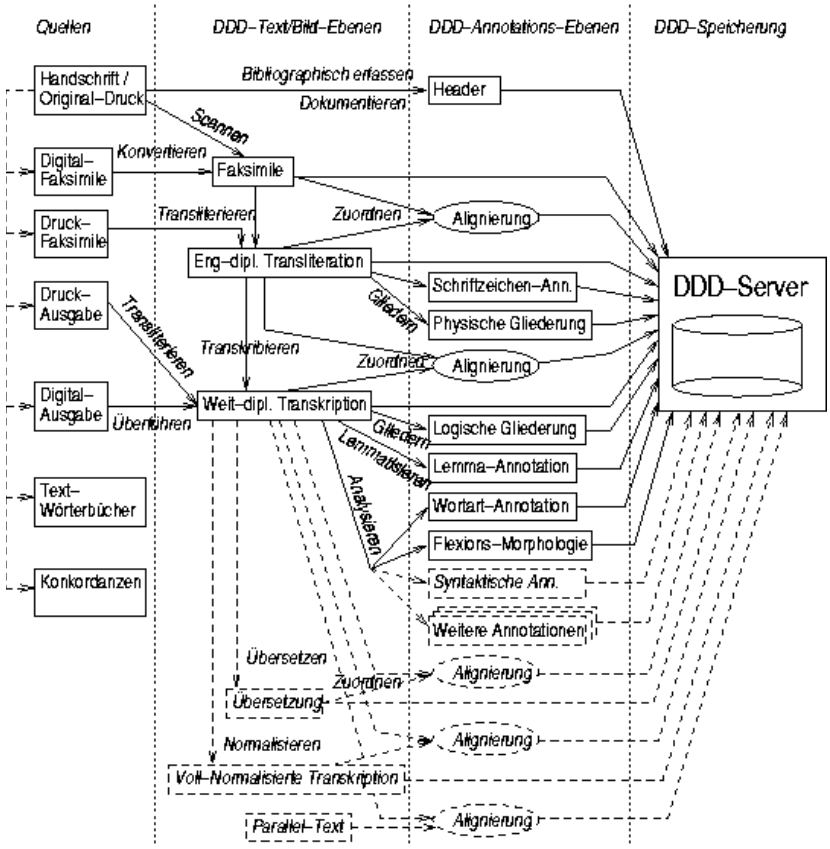

Abbildung 2: Produktionsablauf im DDD-Projekt

Grundsätzlich soll möglichst von den Primärtexten ausgegangen werden, also von Original-Handschriften oder Original-Drucken; beide liegen entweder als Urstück oder in papier- oder filmfaksimilierter Gestalt zur Bearbeitung vor. Zudem gibt es von vielen Original-Texten mehr oder minder handschriftgetreue Druckausgaben, welche zu Rate gezogen werden können (und in einigen Fällen sogar als eigene zu digitalisierende „Primärquelle“ gewertet werden).

Als erstes wird der zu bearbeitende Text in möglichst allen greifbaren primären (Original-Handschriften oder -Drucke) und sekundären Quellen (Papier- oder Film-Faksimiles und Druck- oder Digital--Ausgaben) gesichtet und bei Eignung und Verfügbarkeit in die Arbeit einbezogen; dabei wird es nur selten vorkommen, daß man beim

Digitalisieren eine Originalquelle vorliegen hat; der Regelfall wird sein, daß entweder die digitale Fassung einer Ausgabe vorliegt, die dann noch mit einem Faksimile abzugleichen und dabei in den DDD-Standard zu überführen ist, oder aber daß unter Zuhilfenahme von Druck-Ausgaben von einem Faksimile weg eine Digitalfassung neu erstellt wird. In jedem Falle werden die Digitalfassungen in Einzelfragen noch am „Urstück" (dem nichtreproduzierten Original) entlanggeführt werden müssen.

Setzen wir uns zur Veranschaulichung folgendes Digital-Faksimile als Ausgangspunkt[9]:

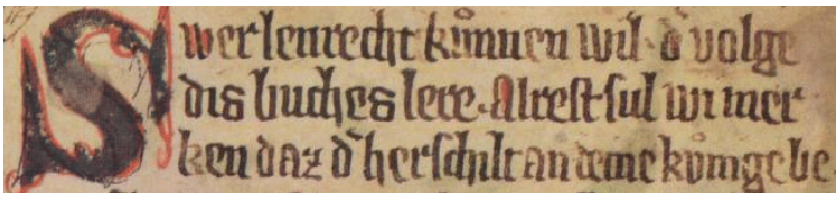

Abbildung 3: Ausschnitt aus einer Sachsenspiegel-Handschrift

Bei der Herstellung eines DDD-Textes nun sollen als erstes im Header die nötigen Angaben ausgefüllt werden (weitgehend nach den Vorgaben der Text Encoding Initiative). Hier zum Beispiel:

```
<DDDCorpus>
        <DDDHeader>
                <title>
                <h.title>Sachsenspiegel</h.title>
                </title>
                <author>
                <h.author>Eike von Repgow</h.author>
                </author>
          ...
        </DDDHeader>
...
</DDDCorpus>
```

[9] Dies sind die ersten Zeilen der Heidelberger Handschrift des Sachsenspiegel. Ein Digital-Faksimile der gesamten Handschrift findet sich unter: <http://digi.ub.uni--heidelberg.de//cpg/cpg164.xml?docname=cpg164&pageid=PAGE0001>

Danach soll ein möglichst gutes Digital-Faksimile – wenn bereits verfügbar – übernommen oder aber selbst eingescannt werden, das dann mit dem Text aligniert werden kann. Im Projektantrag versprechen wir eine seitenweise Alignierung, wir werden aber versuchen, genauer zu sein und möglichst textwortweise zu alignieren.

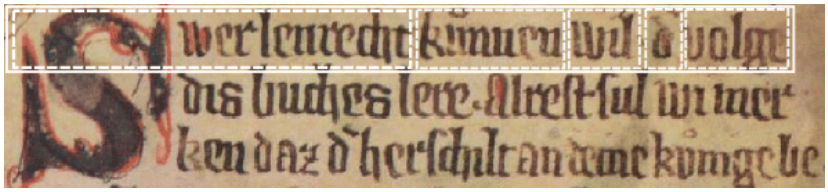

Abbildung 4: Ausschnitt aus dem Sachsenspiegel mit graphischen Markierungen

Dann erfolgt unter Ausnutzung verfügbarer vorhandener Arbeit (etwa bereits elektronisch erfaßter Texte, aber auch bestehender Druck--Ausgaben und Druck-Konkordanzen sowie Druck-Textwörterbücher) eine den DDD-Maßgaben genügende eng-diplomatische Transliteration. Die jeweils urkundengetreueste (idealerweise also die eng-diplomatische) Fassung ist die Referenzfassung für alle weiteren Schritte. DDD nimmt alle Texte, die bis hierher bearbeitet sind, in das sogenannte Erweiterungskorpus auf

1	2	3	4	5	6	7	8	9	10	11	…																	
S	w	e	r	l	e	n	r	e	c	h	t		k	ů	n	n	e	n		w	i	l	·	d̃		v	o	l	g	e

Tabelle 1: Eng-diplomatische Transliteration der ersten Zeile der Sachsenspiegel-Handschrift

Aus der eng-diplomatischen Fassung erstellen wir abstrahierend eine weit-diplomatische Textfassung, welche im Unterschied zu ersterer nicht mehr alle allographischen Feinheiten beibehält, sondern etwa Kürzelstriche oder Schlußbuchstabenformen normalisierend auflöst bzw. vereinheitlicht. Für die sprachwissenschaftlichen Annotationen dürfte sinnvollerweise diese Textfassung zugrundegelegt werden, da sie einerseits nicht mit schreiberischen/ druckerischen Feinheiten des Originals überfrachtet ist, andererseits aber so urkundennah ist, daß alle sprachlich wichtigen Merkmale abfragbar sind.

S	w	e	r	l	e	n	r	e	c	h	t		k	ů	n	n	e	n		w	i	l	·	d	*e*	*r*		v	o	l	g	e

Tabelle 2: Mit Tabelle 1 korrespondierende weit-diplomatische Fassung

Beide Fassungen sind miteinander aligniert:

S	w	e	r	l	e	n	r	e	c	h	t		k	ů	n	n	e	n		w	i	l	·	d̃		v	o	l	g	e
S	w	e	r	l	e	n	r	e	c	h	t		k	ů	n	n	e	n		w	i	l	·	d*er*		v	o	l	g	e

Tabelle 3: Alignierung von eng-diplomatischer und weit-diplomatischer Fassung

Man hat also gewissermaßen zwei Sichten auf den Text (die natürlich miteinander verbunden sind). An jede Fassungen können geeignete positionelle Annotationen andocken. Zweckmäßigerweise an die eng--diplomatische Fassung können Größen wie graphische Struktur, Schreiberhände, Schriftarten, Tintenfarben, Schriftverblassungen oder Pergamentschäden annotiert werden (Tabelle 4), an die weit--diplomatische die logische Struktur, Lemmanamen, flexionsmorphologische Angaben usw. (Tabelle 5).

Ann.-Ebenen	**Einheiten des eng-diplomatischen Textes**																								
	S	w	e	r	l	e	n	r	e	c	h	t		k	ů	n	n	e	n		w	i	l
Typographie	I									lg															
Farbe	r	r																							
graph. Wörter	1													2								3			
graph. Struktur	1. Zeile →																								

Tabelle 4: Mögliche Annotationsebenen der eng-diplomatischen Fassung („I“ steht für „Initiale“, „lg“ für „Ligatur“, „r“ für „rot“)

Ann.-Ebenen	**Einheiten des weit-diplomatischen Textes**			
	S w e r l e n r e c h t	k ů n n e n	w i l
logische Wörter	1	2	3	4
Lemmaname	*swër*	*lê(he)n-rëht*	*künnen*	*wëllen*
Wortart	Pron. 3.Pers.	Substantiv (n.)	Verb	Verb
nhd. Bedeutung	*wenn einer*	*Lehnsrecht*	*verstehen*	*wollen*
Paradigmastelle	Nom.Sg. persönl.	Akk.Sg.	Inf.Präs.	3.Sg.Ind. Präs.
logische Struktur	1. Satz →			

Tabelle 5: Mögliche Annotations-Ebenen der weit-diplomatischen Fassung

Alle Textfassungen und Annotationsebenen sind zeichen- oder zeichenkettenweise miteinander aligniert. Zu jedem Text können dann jederzeit weitere Annotationsebenen (zum Beispiel syntaktische Struktur oder literaturwissenschaftliche Angaben) hinzugefügt werden.

Die Ablage aller Digital-Faksimiles, aller Digitaltextfassungen und aller Zusatzannotationen erfolgt zentral auf dem DDD-Server.

Zusammenfassung

In diesem Papier haben wir die Konzeption eines diachronen Korpus des Deutschen dargestellt, das Texte vom 9. bis zum 19. Jahrhundert enthalten und für möglichst viele textbezogene Wissenschaften zugänglich und nutzbar sein soll.

Ein solches multilinguales und multimodales Korpus braucht zum einen eine Architektur, die das Hinzufügen von Texten und Annotationsebenen erlaubt, zum andern eine weitreichende Standardisierung innerhalb der Annotationsebenen. Wir haben gezeigt, wie das DDD-Projekt (nach heutigem Planungsstand) diese Anforderungen umsetzen wird. Zum Schluß haben wir erläutert, wie ein Text für das Korpus bearbeitet werden soll.

Wir hoffen, mit DDD in naher Zukunft eine wertvolle und langfristig nutzbringende Forschungsressource für die Linguistik, die Philologie

und alle an historischen Texten Interessierten erstellen zu können.

Literatur

Douglas Biber: Representativeness in Corpus Design. In: Literary and Linguistic Computing 8 (1993), S. 243–257

Norman Blake/ Peter Robinson (Hgg.) (1993–1997): The Canterbury Tales Project: Occasional Papers. (Bände I–II). Office for Humanities Communication, Centre for Computing in the Humanities, King's College, London.

Sabine Brants/ Stefanie Dipper/ Silvia Hansen/ Wolfgang Lezius/ George Smith: The TIGER Treebank. In: Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20–21, Sozopol, Bulgaria, 2002.
<http://www.coli.uni-sb.de/~sabine/tigertreebank.pdf> (21.11.2004).

Thomas Burch/ Johannes Fournier/ Kurt Gärtner/ Andrea Rapp (Hgg.): Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2001. Akademie der Wissenschaften und der Literatur, Mainz 2003.

Roberto Busa: Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur, quaeque auspice Paulo VI Summo Pontifice consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa. Stuttgart 1974–1980

Jean Carletta/ Jonathan Kilgour/ Tim O'Donnell/ Stefan Evert/ Holger Voormann: The NITE Object Model Library for handling Structured Linguistic Annotation on Multimodal Data Sets. In: Proceedings of the EACL Workshop on Language Technology and the Semantic Web, 2003
<http://www.ltg.ed.ac.uk/~jeanc/nlpxml2003.final.pdf> (21.11.2004)

Oliver Christ: A modular and flexible architecture for an integrated corpus query system. COMPLEX'94, Budapest 1994 <http://www.ims.--stuttgart.de/CorpusWorkbench/Papers/christ:complex94.ps.gz> (21.11.2004).

Stefanie Dipper/ Lukas Faulstich / Ulf Leser/ Anke Lüdeling: Challenges in Modelling a Richly Annotated Diachronic Corpus of German. In: Proceedings of the Workshop on XML-Based Richly Annotated Corpora. Post-Conference Workshop der LREC 2004. Lissabon 2004 < http://www.deutschdiachrondigital.de/publikationen/dipper_etal_XBRAC04.pdf> (21.11.2004)

Stefan Evert/ Arne Fitschen: Textkorpora. In: Ralf Klabunde et al. (Hgg.) Computerlinguistik und Sprachtechnologie. Eine Einführung. 2. überarbeitete und erweiterte Auflage: Spektrum Akademischer Verlag, Heidelberg 2004, S. 406–413.

Stefan Evert: The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Universität Stuttgart, 2004.

Lukas C. Faulstich / Ulf Leser / Anke Lüdeling: Storing and Querying Medieval Texts in a Database. Technischer Report, Institut für Informatik, Humboldt--Universität zu Berlin, 2004 [in Vorbereitung].

Paul Gévaudan: Klassifikation des lexikalischen Wandels. Semantische, morphologische und stratische Filiation Dissertation, Universität Tübingen 2002. <http://homepages.uni-tuebingen.de/paul.gevaudan/Filiation.pdf> (19.11.2004)

Susan Hockey: Electronic Texts in The Humanities: Principles and Practice: Oxford University Press, New York 2001.

Susan Hockey: Digital Resources in the Humanities: Past, Present and Future: Towards a Universal Digital Library for the Humanities. In: Thomas Burch et al., 2003, Ss. 51–69.

Graeme Kennedy: An Introduction to Corpus Linguistics: Longman, London 1998.

Thomas Klein: Zur Frage der Korpusbildung und zur computergestützten grammatischen Auswertung mittelhochdeutscher Quellen. In: Wegera, Klaus-Peter (Hg.) Mittelhochdeutsche Grammatik als Aufgabe. Zeitschrift für deutsche Philologie 110 (Sonderheft) 1991, Ss. 3–23

Emil Kroymann/ Sebastian Thiebes/ Anke Lüdeling/ Ulf Leser: Eine vergleichende Analyse von historischen und diachronen digitalen Korpora. Technischer Bericht, Institut für Informatik, Humboldt-Universität zu Berlin, 2004.
<http://www.deutschdiachrondigital.de/publikationen/TRHistorischeKorpora.pdf> (21.11.2004).

Merja Kytö: Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts. 3rd edition. Department of English, University of Helsinki 1996.
<http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM> (21.11.2004).

Wolfgang Lezius *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora* Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung 8(4), Stuttgart 2002.
<http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/diss/disslezius.pdf> (21.11.2004)

Chris Manning/ Hinrich Schütze: Foundations of Statistical Natural Language Processing: MIT-Press, Cambridge (MA) 1999.

Tony McEnery/ Andrew Wilson: Corpus Linguistics. Edinburgh: Edinburgh University Press, Edinburgh 2003.

Matti Rissanen/ Merja Kytö/ Minna Pallander: Early English in the Computer Age: Explorations through the Helsinki Corpus: Mouton de Gruyter, Berlin 1993.

Anne Schiller/ Simone Teufel/ Christine Thielen/ Christine Stöckert: Guidelines für das Taggen deutscher Textkorpora mit STTS. IMS Stuttgart und SfS Tübingen 1995.
<http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz > (19.11.2004)

John Sinclair (1996) EAGLES. Preliminary Recommendations on Corpus Typology.
<http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html> (21.11.2004).

Hans-Joachim Solms/ Klaus-Peter Wegera: Das Bonner Frühneuhochdeutschkorpus. Rückblick und Perspektiven. In: Rolf Bergmann (Hg.) Probleme der Textauswahl für einen elektronischen Thesaurus. Stuttgart: Hirzel 1998, S. 22–39

Antonio Zampolli: Past & On-Going Trends in Computational Linguistics: a View from the Instituto die Linguistica Computizionale. In: The ELRA Newsletter, Vol 8(3) (2004), Paris, S. 6–16.

Angegebene Korpora, Digital-Editionen und Ressourcen

- Canterbury Tales Project, <http://www.cta.dmu.ac.uk/projects/ctp/index. html> (21.11.2004)
- CES (Corpus Encoding Standard for XML) <http://www.xml-ces.org/> (21.11.2004)
- Mittelhochdeutsches Wörterbuch < http://www.mhdwb.uni-trier.de/> (21.11.2004)
- NITE (Natural Interactive Tools Engineering) <http://nite.nis.sdu.dk/aboutNite/> (21.11.2004)
- ISO/TC 37/SC 4 <http://www.tc37sc4.org/> (21.11.2004)
- TEI (Text Encoding Initiative) < http://www.tei-c.org/> (21.11.2004)
- STTS (Stuttgart-Tübingen-Tag-Set) < http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html> (21.11.2004)

Anke Lüdeling, Thorwald Poschenrieder und Lukas Faulstich
Korpuslinguistik
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
Unter den Linden 6
D–10099 Berlin
<anke.luedeling@rz.hu-berlin.de>
<http://www.linguistik.hu-berlin.de/korpuslinguistik>