Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas & Frank Wiegand

Berlin-Brandenburgische Akademie der Wissenschaften
Deutsches Textarchiv
Jägerstraße 22/23
D-10117 Berlin
redaktion@deutschestextarchiv.de

Zusammenfassung

Das Deutsche Textarchiv (DTA) macht ein linguistisch annotiertes Volltextkorpus über das Internet frei zugänglich. Die Auswahl beinhaltet sowohl belletristische als auch Fachtexte, die im 17., 18. und 19. Jahrhundert in deutscher Sprache erschienen sind. Das DTA bietet ein vielseitiges Textkorpus für sprachgeschichtliche Forschungen und computerlinguistische Analysen. Durch seine disziplinenübergreifende Zusammensetzung eignet es sich darüber hinaus beispielsweise für fachspezifische, literatur- oder buchwissenschaftliche Untersuchungen. Der Beitrag erläutert die Zusammensetzung des DTA-Korpus, stellt die inhaltlichen und technischen Hintergründe des Projekts vor und gibt einen Ausblick auf die geplante Entwicklung des DTA zu einem aktiven Archiv.

Einleitung

Das Deutsche Textarchiv (DTA, www.deutschestextarchiv.de), ein von der Deutschen Forschungsgemeinschaft (DFG) gefördertes Projekt an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW), stellt ein disziplinenübergreifendes, linguistisch annotiertes Volltextkorpus deutschsprachiger Texte bereit. Grundlage hierfür bilden digitale Faksimiles historischer Druckwerke, die im Zeitraum von ca. 1650 – 1900 erschienen sind. Die Bilddigitalisate und die darauf basierenden elektronischen Volltexte sind über die Website des Projekts öffentlich verfügbar. Die Nutzer des DTA können zwischen einer Bild- und einer Textansicht wählen, wobei letztere in HTML oder einem strukturell reicheren, auf den Empfehlungen der Text Encoding Initiative (TEI P5) beruhenden XML-Format angeboten wird.

In der ersten Projektphase (Juli 2007 bis November 2010) wurden etwa 700 Druckbände der Zeit zwischen 1780 und 1900 im Volltext digitalisiert (dies entspricht einem Umfang von etwa 500 Millionen Zeichen). In seiner zweiten Phase wird das Projekt bis zum Jahr 2014 weitere 650 Texte des 17. und 18. Jahrhunderts im Volltext digitalisieren. Diese werden auf der DTA-Website in regelmäßigen Abständen veröffentlicht. Zum Abschluss von Phase 2 wird das Projekt somit mehr als 1.300 historische Drucke eigendigitalisiert haben und als strukturierte elektronische Volltexte anbieten können.

Zusammensetzung des Korpus

Die Auswahl von Texten für das DTA fand unter sprachwissenschaftlich-lexikographischen Gesichtspunkten statt. In das Korpus wurden Werke aufgenommen, die in der Geschichte der (deutschsprachigen) Literatur oder für die Entwicklung wissenschaftlicher Disziplinen einflussreich waren und die intensiv rezipiert wurden. Neben solchen, als kanonisch geltenden Werken wurden auch einige weniger bekannte Texte berücksichtigt, um die Ausgewogenheit des Korpus zu erhöhen. Die Textauswahl bietet ein großes Spektrum an Genres der literarischen und wissenschaftlichen Produktion.

Zur Zusammenstellung des Korpus wurden im Vorfeld des Projekts zunächst einschlägige Literaturgeschichten und (Fach-)Bibliographien ausgewertet. Parallel dazu wurde das umfangreiche Quellenverzeichnis des *Deutschen Wörterbuchs* von Jacob und Wilhelm Grimm konsultiert.[1] Einzelne Titel, die sich bei der Erstellung bzw. Fortführung des *Deutschen Wörterbuchs* als lexikographisch besonders ergiebig erwiesen haben, wurden der Auswahl hinzugefügt. Ergebnis der Recherchen war eine umfangreiche ‚Basisliste' deutschsprachiger Werke des 17., 18. bzw. 19. Jahrhunderts. Diese Basisliste wurde kommentiert und ergänzt von Mitgliedern der BBAW sowie externen Spezialisten, die sie im Hinblick auf ihr jeweiliges Fachgebiet prüften.

Um den historischen Sprachstand möglichst unverfälscht zu erfassen, wurden die in deutscher Sprache herausgegebenen Erstausgaben der jeweiligen Werke herangezogen. Der Begriff „Erstausgabe" bezieht sich im Zusammenhang mit dem DTA in der Regel auf die *erste gedruckte, selbstständige Publikation eines Textes*. Von diesem Prinzip wurde nur in Einzelfällen abgewichen, etwa dann, wenn eine seltene Erstausgabe eines bestimmten Werks aus konservatorischen Gründen nicht digitalisiert werden konnte (vgl. auch Duntze & Hartwieg, 2010, pp. 63 – 64) oder wenn eine spätere Ausgabe von der Forschung als maßgebliche Fassung des betreffenden Werks angesehen wird – z. B. eine vom Autor überarbeitete und vermehrte Ausgabe oder eine Ausgabe letzter Hand.

Vom gedruckten Buch zum elektronischen Volltext

Das Deutsche Textarchiv arbeitete in der ersten Projektphase vornehmlich mit der Staatsbibliothek zu Berlin, der Herzog August Bibliothek Wolfenbüttel, der Niedersächsischen Staats- und Universitätsbibliothek Göttingen und weiteren, kleineren Bibliotheken zusammen, die Exemplare aus ihren Beständen zur Verfügung stellten. Neben diesen über die erste Projektphase hinaus bestehenden Kooperationen sollen weitere für Phase 2 etabliert werden.

Die Digitalisierung erfolgt in den Räumlichkeiten der Bibliotheken, entweder durch die Bibliotheken selbst oder durch Dienstleister. Über die eigens vom DTA beauftragten Digitalisierungen hinaus profitiert das Projekt von den wachsenden digitalen Sammlungen der Bibliotheken, deren Bilddigitalisate im Rahmen entsprechender Vereinbarungen übernommen werden können. Der im DTA erstellte Volltext steht anschließend auch den Kooperationspartnern zur Verfügung.

[1] Hierbei wurde eng mit der Arbeitsstelle „Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm" zusammengearbeitet, die ebenfalls an der BBAW beheimatet ist.

Die Bilddigitalisate der historischen Drucke bilden die Grundlage für die projektinterne Vorstrukturierung, die mit Hilfe eines vom DTA entwickelten ‚ZoningTools'[2] vorgenommen wird. Damit können Strukturcharakteristika der Seiten, wie beispielsweise Überschriften, Illustrationen oder Marginalien gekennzeichnet werden. Anschließend werden die Texte im sogenannten Double-Keying-Verfahren durch einen Dienstleister erfasst. Dabei werden strukturelle und typographische Merkmale der Vorlage in einem gegenüber dem TEI-Zielformat reduzierten Markup von den abtippenden Personen ausgezeichnet. Die Konvertierung in das Zielformat TEI P5 erfolgt weitestgehend automatisch durch Skripte, die vom DTA entwickelt wurden. Projektmitarbeiter des DTA führen abschließend eine Qualitätskontrolle der Dokumente durch.

Die Bereitstellung der Volltexte im TEI-P5-Format soll deren Nachnutzung erleichtern, beispielsweise als Basis für die Herstellung von textkritischen Ausgaben. Daneben erleichtern die standardisierten Metadaten die Verknüpfung der Texte mit anderen Korpusbeständen.

Bei der computerlinguistischen Aufbereitung, die auf den Texten im strukturierten TEI-P5-Format aufsetzt, werden die Texte in Worteinheiten zerlegt und diese auf ihre jeweilige Grundform zurückgeführt (lemmatisiert). Historische Schreibweisen, die von der heutigen Orthographie abweichen, werden auf die gegenwartssprachliche Variante abgebildet (vgl. Jurish, 2010). Diese Zuordnung ist nicht nur wortbasiert, sondern bezieht durch eine kontextsensitive Disambiguierung auf der Basis eines dynamischen Hidden Markov Models auch noch den engeren Satzkontext mit ein (Jurish, forthcoming). Dadurch wird die Genauigkeit dieser Zuordnung erhöht. Durch diese Analyseschritte ist das Korpus schreibweisentolerant und orthographieübergreifend durchsuchbar. Darüber hinaus wird das DTA-Korpus in einem automatisierten Verfahren hinsichtlich der Wortarten annotiert und für den Substantivbereich mit Thesaurusinformationen verknüpft. Grundlage hierfür bilden der Part-of-Speech Tagger moot (vgl. Jurish, 2003) und die konzeptbasierte lexikalische Begriffshierarchie LexikoNet.[3] Sämtliche linguistische Annotationen geschehen im Standoff-Verfahren. Auf ‚Normalisierungen' des historischen Ausgangstexts wird somit gänzlich verzichtet.

Die Suchmaschine DDC, die für das *Digitale Wörterbuch der deutschen Sprache* (www.dwds.de) entwickelt wurde, ermöglicht neben der gezielten Suche nach einer bestimmten Zeichenkette auch komplexe Suchanfragen sowie die Suche nach flektierten Formen. Die Suche im DTA kann über den gesamten verfügbaren Bestand ausgeführt oder wahlweise auf einzelne Strukturelemente wie Überschriften oder Fußnoten, auf bestimmte typographische Besonderheiten oder einzelne Bände begrenzt werden. Text und Bilddigitalisat wurden vorab in einem automatisierten Verfahren wortweise verknüpft, so dass Treffer von Suchanfragen auch auf dem Digitalisat sichtbar werden.

[2] ZOT – ZoningTool zur Makrostrukturierung von Bilddigitalisaten, vgl. www.deutschestextarchiv.de/documentation/resources/#part_3.

[3] Vgl. www.dwds.de/erschliessung/LexikoNet. In der zweiten Projektphase des DTA soll diese Erschließung durch die Nutzung von Normdaten wie der Gemeinsamen Normdatei (GND) der Deutschen Nationalbibliothek ausgebaut werden.

Ausblick: Das DTA als aktives Archiv

Unter dem Stichwort ‚aktives Archiv' wird das DTA in den kommenden Jahren weiterentwickelt werden. Die Nutzer des DTA sollen die Möglichkeit bekommen, noch intensiver mit dem Datenbestand zu interagieren.

Bereits jetzt ist es möglich, persistente Lesezeichen auf Bildausschnitte der Digitalisate zu setzen, um diese als Belegstellen in einem persönlichen Account zu speichern und mit eigenen Kommentaren zu versehen. Eine vergleichbare Funktion für die Volltexte ist vorgesehen: Markierte Passagen sollen mitsamt der seitengenauen bibliographischen Information im eigenen Bereich gespeichert, annotiert und kommentiert werden können. So bleiben die Fundstellen in dem umfangreichen Korpus überschaubar, und das kooperative Annotieren von Texten in Seminar- und Forschungsgruppen wird dadurch erleichtert. Belegstellen von Text und Bild sollen per E-Mail verschickbar und problemlos exportierbar sein.

Im Bereich der Metadaten bietet das DTA drei verschiedene Schnittstellen an. Literaturmanagementprogramme werden durch das Angebot von COinS (ContextObjects in Spans, http://www.ocoins.info/) unterstützt. Mit dieser Methode zur Einbindung von bibliographischen Metadaten in HTML-Seiten können die Metadaten des DTA automatisch von den weit verbreiteten Programmen Citavi oder Zotero gespeichert werden. Zum Austausch von Autoreninformationen bietet das DTA das relativ junge Format PND-BEACON an. Über diese Schnittstelle können alle im DTA vertretenen Personen zusammen mit der Anzahl ihrer Werke im DTA abgerufen werden. Durch PND-BEACON wird somit eine effiziente Vernetzung mit anderen Projekten möglich, die wie das DTA personenbezogene Daten erfassen. Schließlich bietet das DTA eine OAI-PMH-Schnittstelle an. Dadurch wird der DTA-Bestand demnächst auch über Portale wie Europeana (www.europeana.eu) oder die derzeit im Aufbau befindliche Deutsche Digitale Bibliothek (www.deutsche-digitale-bibliothek.de) recherchierbar sein.

Erweiterung des Korpus

In der zweiten Projektphase soll der aus Eigenmitteln digitalisierte Bestand, das ‚DTA-Basiskorpus', durch Kooperationen mit anderen Projekten angereichert werden, um dem Ziel, das DTA zu einem repräsentativen Querschnittskorpus der deutschen Sprache von 1650 – 1900 auszubauen, näher zu kommen. Hierfür wird derzeit eine umfangreiche Dokumentation des sogenannten DTA-Basisformats erstellt. Dieses enthält die vollständige Liste der TEI-P5-Elemente, die im Projekt verwendet werden, sowie die semantische Beschreibung der Verwendungsweisen dieser Elemente in den Texten des DTA. Zahlreiche Beispiele aus dem DTA-Projektkontext werden diese Dokumentation ergänzen. Darüber hinaus sollen Konvertierungsroutinen zur Verfügung gestellt werden, mit denen Dritte ihre bereits digitalisierten Texte in das DTA-Basisformat überführen können. Die bereits angesprochene Nachnutzung von Digitalisaten aus Bibliotheksbeständen wird die textuelle Anreicherung flankieren.

Literaturverzeichnis

Duntze, O., Hartwieg, U. (2010). Über den Kanon hinaus. Das Deutsche Textarchiv kooperiert mit der Staatsbibliothek zu Berlin. Bibliotheksmagazin H. 1/2010, 62 – 66.

Jurish, B. (2010). Comparing canonicalizations of historical German text. Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON), 72-77, Uppsala, Sweden, 15 July 2010. Retrieved 20 December, 2010, from www.aclweb.org/anthology/W10-2209.

Jurish, B. (2003). A Hybrid Approach to Part-of-Speech Tagging. Final Report, Projekt Kollokationen im Wörterbuch, BBAW, Berlin. Retrieved 20 December, 2010, from http://www.ling.uni-potsdam.de/~moocow/pubs/dwdst-report.pdf.

Jurish, B. (forthcoming). More than words: Using token context to improve canonicalization of historical German. To appear in Journal for Language Technology and Computational Linguistics (JLCL), 2011. See http://www.jlcl.org.