URN: urn:nbn:de:kobv:b4-opus-24424

ALEXANDER GEYKEN,
Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv,

in:
Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011, herausgegeben von Ingelore Hafemann, Berlin 2013, S. 221-234.

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae 4

Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae

4

BERLIN 2013

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Perspektiven einer corpusbasierten historischen Linguistik und Philologie

Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN

2013

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Dieser Band wurde durch die gemeinsame Wissenschaftskonferenz im Akademienprogramm mit Mitteln des Bundes (Bundesministerium für Bildung und Forschung) und des Landes Berlin (Senatsverwaltung für Wirtschaft, Technologie und Forschung) gefördert

URN: urn:nbn:de:kobv:b4-opus-24310

INHALTSVERZEICHNIS

VORWORT

Die internationale Tagung „Perspektiven einer corpusbasierten historischen Linguistik und Philologie“ vom 12. – 13. Dezember 2011 am Akademienvorhaben „Altägyptisches Wörterbuch“ der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) war dem Thema des Aufbaus und der Nutzungsperspektiven elektronischer Textcorpora und Wörterbücher in den historischen Sprachen gewidmet. Die Teilnehmer, Vertreter der Ägyptologie, der Hethitologie, Indogermanistik sowie Referenten aus der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen diskutierten vor allem über die Veränderungen, die mit dem Einsatz elektronischer Erfassungs- und Verarbeitungsprozeduren einhergehen. Vertreter der Computerlinguistik vom „Zentrum Sprache“ der BBAW wurden in die Diskussionen einbezogen. Dort beschäftigt man sich seit Jahren mit dem Aufbau großer elektronischer Textcorpora (DWDS), darunter auch solcher, die historische Texte (DTA) für die elektronische Nutzung ermöglichen.

Die größte Herausforderung dieser neuen elektronischen Corpora und Wörterbücher ist es, sowohl den Methoden und damit den wissenschaftlichen Ansprüchen der traditionellen Philologie und Lexikographie unbedingt verpflichtet zu bleiben als auch neue Gebiete wie die Corpus- und Computerlinguistik für die historischen Sprachen zu öffnen. Die Teilnehmer haben gemeinsam und disziplinenübergreifend die Möglichkeiten und Grenzen der Datenerfassung, ihrer Präsentation und den Nutzen neuer Auswertungsprozeduren diskutiert.

Unter dem ersten Thema „Historische Corpusprojekte – synchron und diachron“ wurden elektronische Corpora vorgestellt und ein intensiver Austausch darüber geführt, welche Datenstrukturen die linguistischen Inhalte in adäquater Weise abbilden. Wichtig war die Frage, auf welche Resonanz diese elektronischen Corpora bei den Nutzern gestoßen sind und welche Erwartungen und Anforderungen aus den verschiedenen Fachdisziplinen an die Projekte herangetragen werden. Der Austausch über Nutzungsperspektiven elektronischer Corpora schloss auch die Diskussion über die Erarbeitung projektübergreifend einsetzbarer Standards der Codierung und Strukturierung historischer Textdaten mit ein. Hinsichtlich einer mittel- und langfristigen Nutzbarkeit sowie einer langfristigen Datensicherheit stehen solche Fragen zunehmend im Focus und einige aktuelle Initiativen dazu wurden vorgestellt. Spezielle technische Aspekte

elektronischer Datenerfassung und automatischer Analyse- und Speicherungsverfahren elektronischer Textdaten konnten am letzten Tag als ein Themenschwerpunkt mit den Programmierern diskutiert werden.

Ein zweiter Schwerpunkt waren konkrete Fragstellungen aus der historischen Lexikographie und diachronen Textanalyse. Für das Ägyptische ist der diachrone Ansatz auf Grund der über viertausendjährigen Textüberlieferung von großer Relevanz. Themen wie historischer und/oder textgattungsspezifischer Wortgebrauch, die Erarbeitung diachroner Wortlisten und Aspekte des kontaktindizierten Sprachwandels konnten disziplinübergreifend zwischen den Ägyptologen und den Kollegen der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen behandelt werden.

Mit dem Abendreferenten Gregory Crane, dem Begründer der „Perseus Digital Library“, wurde ein breites Publikum angesprochen. In seinem Vortrag hat er noch einmal die hohe Relevanz und die neuen Möglichkeiten der Einbeziehung zahlreicher Wissenschaftler und einer interessierten Öffentlichkeit in die Projektarbeit demonstriert, die das Internet auf völlig neue Weise eröffnet hat. Die Herausgeberin ist sehr froh, seinen programmatischen Beitrag zu diesem Thema, dessen schriftliche Form er gemeinsam mit Alison Babeu erarbeitet hat, ebenfalls in diesem Band präsentieren zu können.

Wir danken der Berlin-Brandenburgischen Akademie der Wissenschaften für die umfassende Unterstützung unserer Projektarbeit und ganz speziell der Vorbereitung dieser Konferenz sowie der Möglichkeit, die Akten auf dem E-Doc-Server der Akademie veröffentlichen zu können.

Der Hermann und Elise geborene Heckmann Wentzel-Stiftung sei hiermit ausdrücklich für die unbürokratische und großzügige finanzielle Unterstützung dieser erfolgreichen Tagung gedankt.

Das Akademienvorhaben „Altägyptisches Wörterbuch“ konnte sich als aktives Mitglied des Weiteren auf das „Zentrum Grundlagenforschung Alte Welt“ stützen, dem alle altertumswissenschaftlichen Vorhaben der BBAW angehören. Dem Zentrum ist es zu danken, dass der Abendvortrag von Gregory Crane einem breiteren Publikum dargeboten werden konnte.

Allen Autoren dankt die Herausgeberin für ihre anregenden Diskussionen und die qualitätvollen Beiträge in diesem Band.

Auf eine Gesamtbibliographie wurde verzichtet und die Abkürzungen der in den ägyptologischen Beiträgen erwähnten Zeitschriften und Reihen folgen dem Lexikon der Ägyptologie, herausgegeben von Wolfgang Helck und Wolfhart Westendorf, Band VII: Nachträge, Korrekturen, Indices, Wiesbaden 1992, XIV-XIX.

Ganz besonders sei schließlich Frau Angela Böhme für die gewissenhafte redaktionelle Bearbeitung der Manuskripte gedankt sowie Dr. Simon Schweitzer für seine Hilfe beim Erstellen des Layouts.

Berlin, Mai 2013 Ingelore Hafemann

WEGE ZU EINEM HISTORISCHEN REFERENZKORPUS DES DEUTSCHEN: DAS PROJEKT DEUTSCHES TEXTARCHIV

ALEXANDER GEYKEN

1. Einleitung

Der Nutzen umfassender Referenzkorpora für die Sprachwissenschaft, digitalen Textsammlungen also, die ausgewogen nach Textsorten und hinreichend groß für den Gegenstand der Untersuchungen sind, ist unbestritten: Referenzkorpora dienen unter anderem als Basis für Forschungen zum Wortschatz, zur Wortgeschichte, zur Grammatik der Textorganisation oder zu kontrastiven Studien zum Vergleich Fachwortschatz – normaler Wortschatz (vgl. SINCLAIR 2005, LEMNITZER 2010).

Für die deutsche Gegenwartssprache existieren mit DeReKo (KUPIETZ 2010) und dem DWDS-Kernkorpus (GEYKEN 2007) zwei für die o.g. Fragestellungen geeignete Korpora. Für die älteren Stadien des Neuhochdeutschen (ca. 1650–1900) fällt die Situation derzeit weitaus weniger befriedigend aus[1]. Hierfür gibt es keine hinreichend großen, nach einheitlichen Standards aufbereiteten und übergreifend abfragbaren Korpora, die als Referenzkorpora verwendbar wären und somit eine Grundlage für die o.g. Untersuchungen bilden könnten. Eine Reihe von Ursachen ist hierfür zu nennen, von denen im Folgenden die wichtigsten aufgeführt werden.

(1) Es fehlen bislang einheitliche verwendete Qualitätsstandards für die Erfassungsgenauigkeit auf Zeichenebene und die vorlagengetreue Wiedergabe der Textbasis. Dies gilt sowohl für die zahlreichen Einzelsammlungen, nicht zuletzt aber auch für die großen Textsammlungen Google Bücher, Wikisource, Zeno.org oder Gutenberg.org[2] und Gutenberg-DE. Diese unterscheiden sich von Referenzkorpora nicht nur durch die opportunistische Vorgehensweise bei der Textaufnahme, sondern auch auf der Ebene der

[1] Für das Althochdeutsche (8. Jh. bis ca. 1050) und das Mittelhochdeutsche (ca. 1050-1350) gibt es u.a. die Initiativen von Titus, der Trierer Arbeitsgruppe und die Initiative DeutschDiachronDigital (DDD). Für das Frühneuhochdeutsche wird das seit Herbst 2011 angelaufene Projekt zu einem 300 Texte umfassenden Frühneuhochdeutsch-Corpus im Rahmen von DeutschDiachronDigital (DDD) die Lage verbessern.

[2] Wir beschränken die Diskussion hier auf die deutschsprachigen Anteile dieser Sammlungen.

Meta- und Objektdaten sowie hinsichtlich der erzielten Erfassungsgenauigkeit und der Vorlagentreue von der für ein Referenzkorpus wünschenswerten Genauigkeit. Bei den Metadaten bleibt man beispielsweise bei Google Bücher oder Gutenberg-DE oft im Unklaren, welche Ausgabe dem Volltext zugrunde liegt; bei Zeno.org gibt es zu den Volltexten keine zugehörigen Bild-Digitalisate, was die Nachprüfbarkeit von Transkriptionen ohne einen Gang in die Bibliothek nahezu unmöglich macht. Gutenberg.org oder Wikisource enthalten oftmals modernisierte Transkriptionen und sind somit für manche sprachhistorische Untersuchungen nur von eingeschränktem Wert. Die Transkriptionsgenauigkeit von Google Bücher ist, da sie ausschließlich per OCR entstanden sind, bei historischen Werken ohne gründliche Nachkorrektur als Bestandteil eines Referenzkorpus nicht verwendbar.

(2) Textannotationsstandards, insbesondere die der Text Encoding Initiative (TEI; BURNARD & BAUMAN 2012) sind im wissenschaftlichen Kontext mittlerweile zunehmend verbreitet und stellen für historische Textsammlungen nahezu einen De-Facto-Standard dar[3]. Man sollte daher annehmen, dass dadurch die Austauschbarkeit der Daten als auch die Interoperabilität, also die unmittelbare Einsetzbarkeit TEI-kodierter Daten auf Korpusplattformen, gewährleistet ist. In der Praxis steht dem entgegen, dass für die Textkodierung verschiedene TEI-‚Dialekte' verwendet werden, deren Unterschiede im Allgemeinen so groß sind, dass die Interoperabilität nicht per se gegeben ist. In jüngerer Zeit wurden zwar einige Basisformate geschaffen (s. hierzu Abschnitt 3.2), um die Nutzbarkeit der Texte in verschiedenen Kontexten zu ermöglichen. Diese Basisformate haben bislang jedoch nur eine geringe Verbreitung gefunden.

(3) Bis vor kurzem wurden Korpusprojekte im Wesentlichen als Einzelprojekte durchgeführt, in denen die Textsammlungen projektspezifisch für einen bestimmten Anwendungszweck transkribiert und annotiert wurden. Erst in den letzten Jahren wurde, nicht zuletzt durch die strategischen Schritte der Forschungspolitik in

[3] Dies gilt nicht für die o.g. nicht in wissenschaftlichen Kontexten entstandenen Textsammlungen. Diese liegen entweder in html oder anderen proprietären nicht xml-Formaten vor und müssen erst in ein konsistentes TEI-Format konvertiert werden. Für einen Teil, vorrangig der literarischen Texte aus Zeno.org, ist dies schon geschehen (allerdings mit Informationsverlust), für Wikisource, Gutenberg.org und Gutenberg-DE werden Teile vom DTA konvertiert (www.deutschestextarchiv.de/dtae).

Richtung Open Access, die Weitergabe von Forschungsdaten thematisiert und rechtliche (offene Lizenzen[4]) und technische Rahmenbedingungen (interoperable Annotationsformate) geschaffen. Auf Forschungsebene wurden diese neuen technischen Möglichkeiten aber bislang noch nicht umfassend umgesetzt.

Mit dem Deutschen Textarchiv (DTA) und dessen Verfügbarkeit im Rahmen des großen Infrastrukturverbundes CLARIN-D[5] wird die technische Basis für ein dynamisch erweiterbares historisches Referenzkorpus geschaffen. Im Folgenden soll zunächst das DTA-Korpus als Grundstock für ein Referenzkorpus beschrieben werden (Abschnitt 2). In Abschnitt 3 werden die Anforderungen beschrieben, die sich für den Aufbau einer solchen Korpus-Infrastruktur ergeben. Abschnitt 4 fasst die Ergebnisse zusammen und gibt einen Ausblick auf weitere Arbeiten.

2. Der Grundstock: Das Deutsche Textarchiv

Ziel des von der Deutschen Forschungsgemeinschaft geförderten und an der BBAW beheimateten Projekts Deutsches Textarchiv (DTA) ist es, einen disziplinenübergreifenden Bestand deutschsprachiger Texte aus der Mitte des 17. bis zum Ende des 19. Jahrhunderts nach den Erstausgaben zu digitalisieren und als linguistisch annotiertes Volltextkorpus im Internet bereitzustellen. Um den historischen Sprachstand möglichst genau abzubilden, werden als Vorlage für die Digitalisierung in der Regel die ersten selbstständigen Ausgaben der jeweiligen Werke zugrunde gelegt. Die Volltexterfassung erfolgt möglichst vorlagengetreu und unter Verzicht auf textkritische Eingriffe und Kommentierungen. Hierzu werden die Texte in einem standardisierten Prozess größtenteils manuell (im *Double Keying*-Verfahren) erfasst. Dies ist aufgrund der Textvorlagen, die überwiegend in Fraktur vorliegen, bedeutend zuverlässiger als eine Texterfassung durch OCR (mit anschließender manueller Nachkontrolle).

Hinsichtlich der Entstehungszeit der für das DTA erfassten Texte sowie in Bezug auf die dabei berücksichtigten Textsorten wird eine größtmögliche Ausgewogenheit angestrebt. Derzeit[6] stehen im DTA Werke im Umfang von etwa 247.000 Seiten aus dem Zeitraum von

---

[4] Z.B. mit der Empfehlung, Forschungsdaten unter einer Creative Commons Lizenz zu veröffentlichen.

[5] CLARIN-D: eine web- und zentrenbasierte Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften, www.clarin-d.de.

[6] Stand Dezember 2011.

1780 bis 1900 als elektronische Volltexte und digitale Faksimiles zur Verfügung. In der zweiten Projektphase (Dezember 2010–2013) soll das Textkorpus auf die Zeit bis ca. 1650 ausgeweitet werden. Im Durchschnitt wird täglich ein weiteres Werk digitalisiert und über das Internet bereitgestellt. Mit einem geplanten Umfang von ca. 1.300 Texten des 17.–19. Jahrhunderts (ca. 100 Millionen Textwörtern bzw. 1 Milliarde Zeichen) entsteht mit dem Deutschen Textarchiv ein großes historisches TEI-kodiertes Kernkorpus deutschsprachiger Texte.

Dieses Korpus dient als Basis und Ausgangspunkt für das Referenzkorpus, welches sich aus dem DTA-Korpus und weiteren Texten aus externen Quellen speisen soll.

3. Anforderungen an eine Infrastruktur zum Aufbau eines integrierten Referenzkorpus

In diesem Abschnitt werden die notwendigen Anforderungen formuliert werden, die an eine technische Infrastruktur zum Aufbau eines integrierten historischen Referenzkorpus für das ältere Frühneuhochdeutsche zu stellen sind.

3.1 Aufbau eines Textsorteninventars

Der Aufbau eines Textsorteninventars sowie die Gewichtung der Textsorten untereinander stellt eine wichtige Voraussetzung für die Erstellung eines Referenzkorpus dar. Dies beinhaltet nicht, dass die dem Referenzkorpus zugrunde liegende Textsammlung alle Textsorten entsprechend ihrer Gewichtung enthalten muss. Es bedeutet jedoch, dass alle Texte über ausreichende Metadaten verfügen müssen, insbesondere Datierung, Textsorte, diatopische Merkmale sowie Textumfang. Aus diesen Metadaten lassen sich dann kriteriengestützt Referenzkorpora extrahieren, die den vorher festgelegten Textsorten und ihrer Gewichtung untereinander möglichst optimal Rechnung tragen. Ein entsprechendes Optimierungsverfahren wurde beim Auswahlprozess des DWDS-Kernkorpus aus einer etwa drei Mal so großen Kernkorpusbasis angewendet (GEYKEN 2007). Als Ergebnis entstand ein optimal nach Textsorten ausgewogenes Textkorpus mit einer Größe von 100 Millionen laufenden Textwörtern.

Neben der Auswahl der Textsorten und deren Gewichtung zueinander müssen auch für die eigentliche Textauswahl Kriterien bestimmt werden. Für das DTA, welches den Grundstock des Referenzkorpus bilden soll, fand die Textauswahl 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.

Die Verteilung der Texte soll hinsichtlich der unterschiedlichen Disziplinen und Textsorten möglichst ausgewogen sein. Die nachstehenden Diagramme 1 und 2 zeigen die diesbezügliche Verteilung der DTA-Texte auf Basis der jeweiligen Anzahl der Titel.

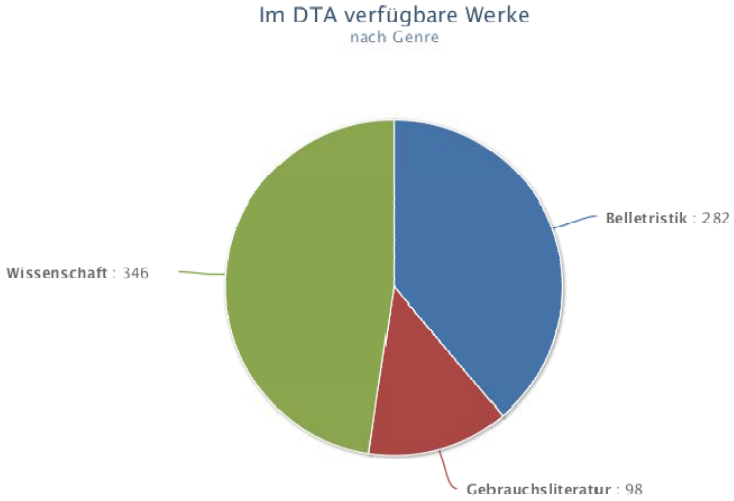

Abb. 1: Textsorteninventar und Verteilung für des DTA (Zeitraum 1780-1900)

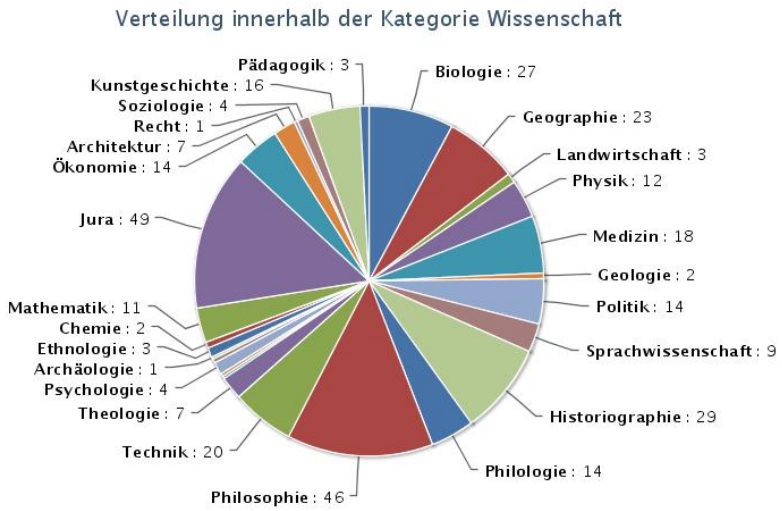

Abb. 2: Verteilung innerhalb der Kategorie Wissenschaft (Zeitraum 1780-1900)

3.2 Einheitlichkeit der Formate

Wie in Abschnitt 1 bereits erwähnt, ist die TEI auf einem guten Wege, ein De-Facto-Standard für die Annotation historischer Texte zu werden. Aufgrund ihrer großen Flexibilität ist die TEI jedoch nicht per se interoperabel (z.B. UNSWORTH 2011). Interoperabilität ist jedoch ein entscheidender Eckpfeiler für eine Korpusinfrastruktur, die auf verteilten Ressourcen basiert. Auf Metadatenebene ermöglicht Interoperabilität, dass Texte einheitlich verarbeitet werden und über große Datenbanken verfügbar gemacht werden können. Auf der Ebene der Text- bzw. Objektdaten erleichtert die einheitliche Annotation die Auswertung der Texte für weitere Analysen, wie z.B. Text Mining oder die tiefere Annotation des Texts mit Diskursinformationen. Darüber hinaus lassen sich einheitlich formatierte Texte auch indizieren, ohne dass dafür Konvertierungen oder manuelle Arbeitsschritte notwendig wären. Für die Abfrage hat dies den Vorteil, dass auch semantische Suchabfragen, sofern sie vorher annotiert wurden, konsistent abfragbar sind. Beispielsweise könnten aus großen Textsammlungen nur dann alle Abschnitte vom Typ „Brief“ mit einer Suchabfrage extrahiert werden, wenn diese vorher auch einheitlich ausgezeichnet wurden. Schließlich können einheitlich kodierte Meta- und Objektdaten in großen Repositorien gespeichert werden und damit nachhaltig vorgehalten werden.

Wie lässt sich die Interoperabilität von Korpusdaten gewährleisten? Die TEI selbst ist sich dessen bewusst, dass die TEI als solche zu unspezifisch für einen interoperablen Einsatz ist und empfiehlt daher die Erstellung von geeigneten Untermengen von TEI-P5 (BURNARD & BAUMANN 2012). Die TEI selbst liefert auch Vorschläge wie TEI Tite (TROLARD 2011), TEI Lite (BURNARD & SPERBERG-MCQUEEN 2006) oder Best Practices for TEI in Libraries (TEI SIG on Libraries 2011). Darüber hinaus gibt es projektspezifische Anpassungen wie beispielsweise TEI Analytics (UNSWORTH 2011; PYTLIK ZILLIG *et al.* 2009), IDS-XCES (Institut für deutsche Sprache, Mannheim) oder Textgrid's Baseline Encoding for Text Data in TEI P5 (Textgrid 2007–2009). Eine Gemeinsamkeit dieser Formate ist, dass sie nur eine gegenüber der gesamten TEI-P5 reduzierte Zahl von Elementen verwenden, um eine bessere Kontrolle über Annotation der Dokumente zu erhalten. Solche Formate, technisch als XML-Schemata beschrieben, sollten genügend ausdrucksstark sein, um eine Basisstrukturierung von Texten zu ermöglichen, die wiederum als Ausgangspunkt für tiefere projektspezifische Annotationen dienen kann.

Das Basisformat des Projekts DTA[7] (fortan DTA-BF) ist ebenso wie die anderen oben genannten Formate eine Untermenge des TEI-P5. Ziel des DTA-BF ist es, für die heterogenen Anforderungen des DTA-Projekts – da Texte aus unterschiedlichen Zeiträumen, Orten und Textsorten aufgenommen werden – eine weitestmöglich eindeutige Kodierungszuordnung für die unstrittigen strukturellen Aspekte bereitzustellen. Insbesondere soll durch das DTA-BF sichergestellt werden, dass semantisch gleiche Phänomene strukturell gleich und damit eindeutig annotiert werden. Ziel des DTA-BF ist somit, die Interoperabilität von Texten auf der Ebene der Metadaten und der Textstrukturierungsebene zu gewährleisten und damit eine übergreifende Einheitlichkeit der Abfragemöglichkeiten sicherzustellen. Ein wichtiger Unterschied gegenüber den oben genannten Formaten besteht vor allem darin, dass das DTA-BF nicht versucht, eine möglichst große Anzahl verschiedener Textkollektionen durch die Bildung der Vereinigungsmenge über all diese Formate zu gewährleisten, wie dies beispielsweise bei TEI-Analytics geschieht. Vielmehr sollen bei der Integration neuer Texte im Vorfeld Redundanzen vermieden werden. Zugleich werden möglichst viele Informationen aus den externen Texten bewahrt, statt einen Teil dieser nach dem Prinzip des ‚kleinsten gemeinsamen Nenners' zu ignorieren. Dies

[7] www.deutschestextarchiv.de/doku/basisformat.

geschieht dadurch, dass neue Elemente auf bereits bestehende Elemente oder Attribut-Wert-Paare abgebildet werden. Somit soll die Anzahl der verschiedenen Elemente auf eine möglichst klein bleiben. Ähnliches gilt für die Attribut-Wert-Paare des DTA-BF. Diese werden durch eine vorgegebene Liste von Werten beschrieben. Bei der Auswertung ist somit gewährleistet, dass es zu keinem „unkontrollierten Wachstum" verschiedener Bezeichnungen für gleiche semantische Inhalte kommt.

Bei dem DTA-BF handelt es sich um ein flexibles Format: das bedeutet, dass Annotationen eines Texts im DTA-BF in unterschiedlicher Tiefe vorgenommen werden können, die durch sogenannte Levels voneinander unterschieden werden. Dies stimmt auch mit den Empfehlungen der TEI überein, die für die Kodierung von Korpora vier Ebenen der Annotation vorschlägt: obligatorische, empfohlene, optionale und verbotene Elemente[8].

3.3 Qualitätssicherung

Um die einheitliche Transkriptions- und Annotationsqualität zu gewährleisten, ist es notwendig, dass die Infrastruktur eine Qualitätssicherungsumgebung enthält, in der die importierten Textquellen evaluiert und gegebenenfalls korrigiert werden können. Die Qualitätskontrolle findet sowohl formativ (im Zuge der Auswahl des geeigneten Exemplars und durch Vorannotation) wie auch summativ statt. Summative Plattformen für das verteilte Korrigieren existieren für die großen Textsammlungen wie Wikisource, Gutenberg-DE[9] oder Gutenberg.org, sie sind jedoch bisher nicht für die Korrektur von TEI-Dokumenten implementiert worden. Für das DTA wurde daher eine solche Umgebung entwickelt: DTAQ[10]. DTAQ ist web-basiert und ermöglicht die verteilte Prüfung (und bislang in eingeschränktem Maße auch die Korrektur) von TEI-Dokumenten. In DTAQ können Strukturierungs- und Erfassungsfehler, aus der Vorlage übernommene Druckfehler sowie im Erfassungsprozess unterlaufene Transkriptionsfehler gemeldet, ggf. kommentiert und nachvollziehbar behoben werden. Für die Textkontrolle haben Nutzer die Möglichkeit, Texte seitenweise in der Gegenüberstellung von Text und Bild anzusehen und Fehler (Transkriptionsfehler, Druckfehler, Strukturierungs- und

[8] www.teic.org/release/doc/tei-p5-doc/en/html/CC.html.

[9] Z.B. www.gaga.net.

[10] www.deutschestextarchiv.de/dtaq/about.

Darstellungsfehler) zu melden. Verschiedene Textansichten sind verfügbar:

- die originale XML/TEI-Fassung,
- eine HTML-Darstellung,
- eine reine Textansicht,
- die Transkription nach automatischer linguistischer Analyse.

DTAQ ist sein Juni 2011 im Einsatz. Seither wurden etwa 17.000 Textseiten vollständig Korrektur gelesen und insgesamt mehr als 30.000 Fehler gemeldet. Derzeit sind knapp 100 Nutzer in DTAQ angemeldet.[11] Spezialanwendungen erlauben die Fokussierung auf bestimmte Phänomene (typischerweise fehlerhafte Zeichenketten, Textmaterial nicht-lateinischer Alphabete etc.) sowie auf wünschenswerte Erweiterungen des Korpus (z.B. die Transkription von Formeln, deren Vorhandensein bisher nur durch ein leeres <formula/>-Element angedeutet wurde, mithilfe eines speziellen Formel-Editors).

3.4 Übergreifende Abfragbarkeit

Für gegenwartssprachliche Texte ist eine übergreifende wortformen- bzw. lemmabasierte Volltextsuche mittlerweile eine Standardanwendung, da für die heutige Orthographie umfassende morphologische Analyseprogramme vorliegen. Damit kann beispielsweise die Suche nach der Wortform „Kleid“ auch Treffer für alle flektierten Formen des morphologischen Paradigmas „Kleids“, „Kleider“, „Kleidern“ etc. zurückliefern.

Da historische Texte, wenn sie gemäß ihrer Originalausgaben transkribiert sind, in keiner standardisierten Rechtschreibung vorliegen (die erste Normierung der deutschen Rechtschreibung fand erst 1902 statt), ist die übergreifende Abfragbarkeit nicht durch Standardprogramme zu leisten. Für das Deutsche gibt es derzeit zwei Verfahren, die beide darauf basieren, historische Formen, die von der heutigen Orthographie abweichen, auf die zugehörige gegenwartssprachliche Form abzubilden (GOTSCHAREK *et al.* 2009, JURISH 2010, JURISH 2011). Mit diesen Verfahren ist es möglich, die orthographisch uneinheitlichen Texte zu durchsuchen, indem die orthographisch gültige Form als Eingabewort verwendet wird. Beispielsweise würde eine Suche nach der heutigen Form „Kleid“ auch die dazugehörigen historischen graphematischen Varianten für „Kleidt“, „Kleidts“ „Kleydt“, „Cleyd“, „Cleit“ etc. finden. Durch diese Analyse-

[11] Stand August 2012.

schritte ist das Korpus somit schreibweisentolerant und orthographieübergreifend durchsuchbar.

3.5 Nachnutzbarkeit der Texte

Die Nachnutzbarkeit von Texten hat eine rechtliche und eine technische Ebene. Auf der rechtlichen Ebene müssen Texte, wenn sie durch Dritte für Forschungszwecke nutzbar sein sollen, mit einer offenen Lizenz versehen sein, damit Analysen oder Zusatzannotationen zu einem Text wieder zusammen mit den Basistexten veröffentlicht werden können. Mittels einer Creative Commons Lizenz ist dies möglich. Texte können damit nicht nur zum Download angeboten werden, sondern auch in andere Repositorien eingebettet und dort für weitere Nutzungsformen zur Verfügung stehen. Im Projekt DTA beispielsweise stehen alle vom DTA produzierten Texte unter einer CC-BY-NC Lizenz. Sie sind somit unter Namensnennung des DTA für nichtkommerzielle Zwecke in beliebigen Forschungskontexten nachnutzbar.

Auf technischer Ebene müssen verschiedene Maßnahmen bezüglich Zugriff, Objektpersistenz und Formatstandardisierung getroffen werden. Zunächst einmal sollten die Dokumente autonom in Repositorien vorgehalten werden und nicht etwa in Datenbanken gekapselt. Des Weiteren sollten Texte besteht mit persistenten Identifizierern (PID) ausgestattet sein, damit sie von Dritten verlässlich referenziert werden können. Die Referenzierungsgenauigkeit sollte dabei mindestens auf Dokumentebene, besser aber auf Seiten- bzw. Zeilenebene erfolgen. Drittens sollten Meta- und Objektdaten gemäß transparenter und weit verbreiteter Standards beschrieben sein. Verwendet man beispielsweise Dublin Core oder klar definierte TEI-Header, können die Metadaten über eine OAI-PMH-Schnittstelle verfügbar gemacht werden und stehen somit für viele Formen der Nachnutzung, insbesondere zur Verzeichnung und Kontextualisierung in den großen Metadatenkatalogen, zur Verfügung. In Abschnitt 3.2 wurde bereits darauf eingegangen, dass Objektdaten, die in einem TEI-Basisformat (Beispielsweise dem DTA-BF) vorliegen, auch interoperabel einsetzbar und somit auch leichter nachnutzbar in anderen Kontexten sind.

3.6 Infrastruktur für die Anreicherung der Textbasis

Die Korpusinfrastruktur muss offen für externe Beiträger gestaltet werden, die im Rahmen ihrer Forschungsarbeiten Texte des späten 16. bis frühen 20. Jahrhunderts bearbeiten und/oder digitalisieren. Dies geschieht zum wechselseitigen Nutzen: Die Beiträger erreichen

durch die Verbreitung ihrer Daten im DTA eine höhere Sichtbarkeit, als wenn sie diese nur privat über die eigene Website oder den Universitätsserver zur Verfügung stellen. Insbesondere bietet das DTA externen Beiträgern folgende Möglichkeiten:

- eine Text-/Bild-Ansicht und eine Leseansicht, jeweils mit einer ausführlichen Präsentation der Metadaten, im DTA sowie – durch Einbettung eines Inlineframes (<iframe>) – auf ihrer eigenen Internetseite,
- einen eigenen Suchindex für die Textkollektion,
- die Korrekturmöglichkeit vor der Veröffentlichung durch die Korrekturplattform DTAQ (s. Abschnitt 3.3), bei der die Texte von einer wissenschaftlichen Nutzergemeinschaft evaluiert werden,
- standardisierte Schnittstellen (OAI-PMH) zum Austausch von Metadaten.

Das DTA kann auf solche Weise das DTA-Kernkorpus fortlaufend durch Primärtexte ergänzt werden, wodurch die Vielfalt und Umfang weiter erhöht wird. Darüber hinaus werden so vergleichende linguistische Untersuchungen zwischen den angelagerten Spezialkorpora und dem DTA-Kernkorpus ermöglicht.

Kandidaten für Ergänzungen zum DTA zeichnen sich durch ihre hohe Wirkungsmächtigkeit aus oder stellen Schlüsseltexte innerhalb wichtiger Diskurse dar. Für die Integration der Texte in das DTA ist die strukturelle Aufbereitung entsprechend dem DTA-Basisformat vonnöten. Das DTA oXygen-Framework unterstützt die Erarbeitung von Texten entsprechend dem DTA-Basisformat. Ziel ist ein hinsichtlich der Transkription und der Annotation homogenes Korpus, dessen Texte uneingeschränkt miteinander interoperabel sind.

Derzeit unterhält das DTA in Kooperation mit 10 Projekten, deren Texte sukzessive über das DTA verfügbar gemacht werden[12] und bemüht sich zudem um die Integration einzelner, qualitativ hochwertiger Textzeugen aus den o.g. kleinen und großen Textsammlungen.

3.7 Aktives Archiv, d.h. lebende elektronische Texte

Ein grundsätzlicher Unterschied zwischen gedruckten Publikationen und elektronische Texten besteht in dem unterschiedlichen Lebenszyklus: gedruckte Publikationen sind statisch, Korrekturen werden allenfalls über eine zweite Auflage kenntlich gemacht. Im Unterschied dazu sind digitale Publikationen einer dynamischen Verände-

[12] www.deutschestextarchiv.de/dtae.

rung unterworfen. Transkriptions- oder Druckfehler können jederzeit vermerkt werden und führen durch die Vergabe einer neuen PID zu einer neuen Version des Dokuments. Wichtiger ist die Tatsache, dass auch die Annotationen des Texts stets verfeinert werden können: beispielsweise können im Laufe der Lektüre oder der Arbeit mit dem elektronischen Text Eigennamen oder Zitate in der Annotationsumgebung von DTAQ markiert und wieder in das Archiv zurückgespielt werden. Notwendig für den Aufbau eines im skizzierten Sinne aktiven Archivs ist somit die Schaffung einer Annotationsumgebung, in der die Annotationen zusammen mit dem Basistext (und dessen Versionen) verwaltet werden können.

4. Zusammenfassung und Ausblick

In diesem Beitrag wurde das Korpus des Deutschen Textarchivs als Basis für ein dynamisch erweiterbares historisches Referenzkorpus vorgestellt. Es wurden sieben Anforderungen für eine Korpus-Infrastruktur benannt, die dazu dienen sollen, in systematischer Weise weitere Texte für die historische Korpusforschung nutzbar zu machen. Dabei wurden rechtliche (OpenAccess) und technische (Standardisierung der Formate) Eckpfeiler benannt.

Notwendige Voraussetzung für den Erfolg ist jedoch die Bereitschaft der Beiträger, sich aktiv an einem solchen Vorhaben zu beteiligen – sei es durch das Bereitstellen eigener Daten oder über die kollaborative Arbeit mit und an bereitgestellten Texten. Dabei darf die vorgeschlagene Infrastruktur nicht als Einbahnstraße erscheinen, sondern muss den Beiträgern glaubhaft vermitteln, dass die in die Infrastruktur hineingegebenen Texte genauso wie alle anderen Texte der Infrastruktur wieder nachgenutzt und in andere Analyse- und Präsentationsumgebungen eingespeist werden können. Mit seiner offenen Lizenzpolitik (CC-BY-NC), die jedem Beiträger genau diese Nachnutzungen ermöglicht, leistet das DTA hierfür einen Beitrag. Ein Projekt alleine ist zu klein, um die verschiedenen potentiellen Beiträger tatsächlich beraten zu können. Die Schaffung einer örtlichen und über verschiedene Institutionen verteilten Arbeitsgruppe ist daher ein nächstes Ziel. Diese soll dazu beitragen, die Sogwirkung zu entfalten, die dafür notwendig ist, dass mehr Beiträger für das Referenzkorpus gewonnen werden können. Eine solche Arbeitsgruppe, bestehend aus dem DTA, der HAB Wolfenbüttel, dem IDS Mannheim und der Universität Gießen wurde im Rahmen des CLARIN-D Projekts im Juni 2012 von der Facharbeitsgruppe Deutsche Philologie gegründet (CLARIN-D Kurationsprojekt).

5. BIBLIOGRAPHIE

BURNARD, L. & S. BAUMAN, 2012: *P5: Guidelines for Electronic Text Encoding and Interchange, Version 2.1.0, June 17th, 2012*. http://www.tei-c.org/release/doc/tei-p5-doc/en/html.

BURNARD, L. & SPERBERG-MCQUEEN, C. M., 2006: *TEI Lite: Encoding for Interchange: an introduction to the TEI – Revised for TEI P5 release*. http://www.tei-c.org/Vault/P5/2.1.0/doc/tei-p5-exemplars/html/teilite.doc.html.

CLARIN-D Kurationsprojekt, 2012: »Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur«, Vorhabensbeschreibung für ein Kurationsprojekt der F-AG 1 »Deutsche Philologie«. http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-1.

GEYKEN, A., 2007: The DWDS corpus: A reference corpus for the German language of the 20th century, in: FELLBAUM, CHR. (ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*, London, 23-41.

GEYKEN, A. *et al.*, 2012: The DTA 'base format': A TEI-subset for the compilation of Interoperable Corpora. To appear in: *Proceedings of Konvens 2012*.

GEYKEN, A. *et al.*, 2012: TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv, in: *Jahrbuch für Computerphilologie*, online-Version vom 05.08.2012. http://www.computerphilologie.de/jg09/geykenetal.html.

GEYKEN, A. *et al.* (Panel): Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of Publication of Digitized Historical Prints. Panel DH2012, Hamburg (Abstract und Video Lecture). http://www.dh2012.uni-hamburg.de.

GOTSCHAREK, A. *et al.*, 2009: Enabling information retrieval on historical document collections: the role of matching procedures and special lexica, in: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, New York, 69–76.

HAAF, S. *et al.*, 2012: Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-

annotated historical text, in: *Journal of the Text Encoding Initiative* 4 (Forthcoming Paper).

JURISH, B., 2010: More than Words: Using Token Context to Improve Canonicalization of Historical German, in: *Journal for Language Technology and Computational Linguistics (JLCL)*, vol. 25/1, 23–39.

JURISH, B., 2011: *Finite-state canonicalization techniques for historical German* (URN: urn:nbn:de:kobv:517-opus-55789) (URL: http://opus.kobv.de/ubp/volltexte/2012/5578/).

KUPIETZ, M. *et al.*, 2010: The German Reference Corpus DeReKo: A primordial sample for linguistic research, in: CALZOLARI, N. *et al.* (eds.), *Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010)*, 1848-1854.

LEMNITZER, L. & H. ZINSMEISTER, 2010: *Korpuslinguistik. Eine Einführung*, 2. Aufl., Tübingen.

PYTLIK ZILLIG, B. L., 2009: TEI Analytics: converting documents into a TEI format for cross-collection text analysis, in: *Literary and Linguistic Computing*, 24 (2), 187-192.

SINCLAIR, J., 2005: Corpus and Text – Basic Principles, in: WYNNE, M. (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, 1-16.

TROLARD, P., 2011: *TEI Tite – A recommendation for off-site text encoding. Version 1.1, September 2011*. www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html.

UNSWORTH, J., 2011: Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI, in: *Journal of the Text Encoding Initiative* 1. http://jtei.revues.org/215 (accessed June 24th, 2012).