Alexander Geyken, Matthias Boenig, Susanne Haaf,
Bryan Jurish, Christian Thomas und Frank Wiegand

10 Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN

Abstract: Im Zentrum dieses Beitrags steht das *Deutsche Textarchiv*, eine Plattform zum Korpusaufbau und zur Korpusanalyse, die im Kontext aller geistes- und sozialwissenschaftlichen Disziplinen mit historischen Fragestellungen nutzbar ist. Das DTA, das am Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) angesiedelt ist, wurde von 2007 bis 2016 von der Deutschen Forschungsgemeinschaft (DFG) gefördert und bildet mittlerweile eine wesentliche Komponente der Forschungsdateninfrastruktur des deutschen Teils von CLARIN. Der Beitrag präsentiert das DTA als webbasierte Forschungsplattform sowohl für die Erstellung und die Kuration von Korpustexten als auch für die Korpusanalyse und verortet es innerhalb der (digitalen) Geisteswissenschaften.

Keywords: Annotation, Historische Fragestellungen, Qualitätssicherung, Sprachkorpora, XML

1 Einführung

Ziel des *Deutschen Textarchivs* (DTA)[1] ist die Erstellung eines disziplinen- und gattungsübergreifenden Grundbestands deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis etwa 1900. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern der BBAW kommentierten und ergänzten, umfangreichen Bibliographie. Aus dieser wurde von der DTA-Projektgruppe ein nach Textsorten und Disziplinen ausgewogenes Textkorpus zusammengestellt, das als Grundlage für ein Referenzkorpus zur Entwicklung der neuhoch-

1 http://www.deutschestextarchiv.de (letzter Zugriff: 6. 11. 2017).

Alexander Geyken, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, Frank Wiegand, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, D-10117 Berlin, E-Mail: dta@bbaw.de

https://doi.org/10.1515/9783110538663-011

deutschen Sprache dient. Um den historischen Sprachstand möglichst genau abzubilden, wurden als Vorlage für die Digitalisierung in der Regel die Erstausgaben der Werke zugrunde gelegt. Das nach diesen Kriterien zusammengestellte DTA-Kernkorpus wird kontinuierlich erweitert. Derzeit umfasst es etwa 1.500 Werke mit einem Umfang von etwa 120 Millionen Textwörtern (Stand April 2017).

Neben seiner Funktion als Korpusaufbauprojekt wurde das DTA von Beginn an auch als aktives Archiv konzipiert. Es soll Ort der Anlagerung weiterer Korpora sein. Hierzu wurden zunächst Qualitätskriterien bezüglich der Metadaten, der Auszeichnungstiefe der Textstrukturen sowie der Genauigkeit des Volltexts festgelegt. Des Weiteren wurde aus der Vielzahl der verschiedenen Texte ein für viele Kontexte nutzbares Strukturformat entwickelt, das neben seiner Funktion als Austauschformat für verschiedene Korpora die Interoperabilität für so verschiedene Anwendungsfälle wie die Korpusanzeige, die Volltextsuche und das Textmining gewährleistet. Mit dem DTA-Basisformat (DTABf) liegt ein solches Format vor, welches mit dem XML/TEI-P5-Standard vollständig kompatibel ist und das mittlerweile auch eine weit über das DTA hinausgehende Verbreitung gefunden hat (dazu mehr in Abschnitt 2.1). Zur Qualitätssicherung der Volltexte und der Strukturdaten wurde darüber hinaus mit DTAQ eine webbasierte Plattform entwickelt, die das verteilte Korrekturlesen und Korrigieren von Texten erlaubt. Hierzu wurden flexible Möglichkeiten zum Textimport aus unterschiedlichen Formaten und eine Text-Bild-Ansicht geschaffen sowie ein Editor in die Plattform integriert, mit dem Texte ohne zusätzlich zu installierende Software bearbeitet werden können (siehe dazu Abschnitt 2.2–4). Am Ende des Korrekturprozesses steht die Veröffentlichung auf der DTA-Webseite, wo die Werke über eine Text-Bild-Ansicht zugänglich sowie an verschiedene Analysewerkzeuge angebunden sind. Letztere führten dazu, dass sich das DTA mit DTAQ von einer Korrektur- und Veröffentlichungsplattform zu einer Forschungsplattform entwickelte. Mit CAB (*Cascaded Analysis Broker*; vgl. dazu Jurish 2012), einem Werkzeug zur Normalisierung historischer Schreibweisen, wird eine schreibweisentolerante Volltextsuche über alle Texte des DTA bereitgestellt. Mit der Integration von GermaNet (Hamp & Feldweg 1997; Henrich & Hinrichs 2010), einer lexikalischen Ressource, die Substantive, Verben und Adjektive nach Bedeutungsähnlichkeit in SynSets zusammenfasst, wird darüber hinaus auch die Volltextsuche nach semantischen Kategorien ermöglicht. Ferner stehen eine Reihe von lexikometrischen Analysewerkzeugen zur Verfügung, insbesondere zu zeitlichen Verläufen von Wortfrequenzen, zu diachronen Kollokationen sowie eine auf den Voyant-Tools basierende quantitative Textanalyse (siehe hierzu Abschnitt 3).

Die Integration von Texten in das DTA stellt keine Einbahnstraße dar: Alle Texte des DTA stehen unter einer offenen Lizenz und können damit ohne

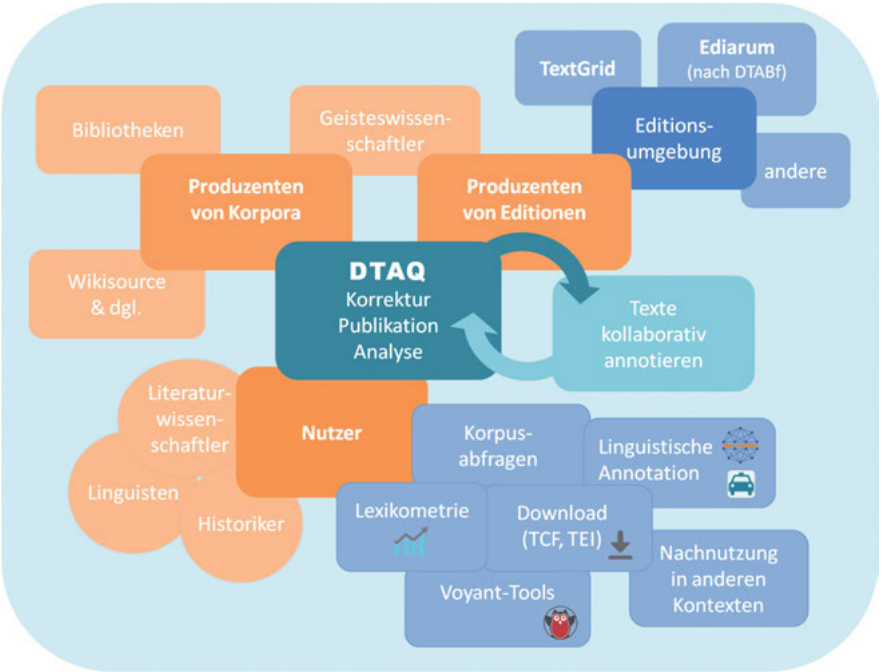

Abb. 10.1: Die Forschungsumgebung DTA.

Weiteres als Ganzes in wissenschaftlichen Kontexten nachgenutzt werden. Aufgrund der durch die Kodierung in DTABf gewährleisteten Interoperabilität können alle Texte des DTA darüber hinaus einfach in verschiedene Formate konvertiert werden (siehe Abschnitt 2.1). Seit 2014 ist das DTA fest in die CLARIN-Infrastruktur eingebunden. Die Möglichkeiten der Kooperation und der Nachnutzungen konnten dadurch weiter ausgebaut werden (siehe Abschnitt 4.2). Das DTA hat sich seitdem zu einer vollwertigen Forschungsplattform entwickelt, zu der Nutzerinnen und Nutzer entweder als Korpusproduzenten beitragen oder die sie als Plattform für die Textanalyse verwenden können. Abbildung 10.1 fasst die verschiedenen Komponenten zusammen, in deren Zentrum DTAQ als Korrektur-, Publikations- und Analyseplattform steht. Auf der einen Seite befinden sich die verschiedenen Korpusproduzenten (Geistes- und Sozialwissenschaftler und -wissenschaftlerinnen, Bibliotheken und außerakademische Initiativen wie beispielsweise Wikisource), auf der anderen Seite stehen Editionsumgebungen und Produzenten von Editionen. Die ‚klassische' Nutzung von DTAQ besteht in der kollaborativen Annotierung von Texten. Alle Texte des DTA können jederzeit korrigiert und annotiert werden, und die stets aktuelle Fassung kann aus der Plattform exportiert werden. Die vierte Kompo-

nente stellt schließlich die Analyse dar, mit den bereits erwähnten Werkzeugen CAB und GermaNet zur linguistischen Annotation, den verschiedenen Analysewerkzeugen und den Exportformaten zur flexiblen Nachnutzung in anderen Kontexten.

2 Ressourcenaufbau und Annotation

2.1 Datengrundlage und Annotationsformat: DTA-Basisformat (DTABf)

Um den Ansprüchen des DTA für eine möglichst vorlagengetreue Transkription historischer Quellen gerecht zu werden und gleichzeitig die Erfassung detailreicher Metadaten und umfangreicher Annotationen logischer und layoutbezogener Strukturen zu ermöglichen, wurde das DTA-Basisformat (DTABf) – ein auf den P5-Richtlinien der *Text Encoding Initiative* (TEI) basierendes XML-Format[2] – entworfen. Das DTABf stellt eine echte Teilmenge der von der TEI vorgegebenen Richtlinien zur Kodierung von Textdokumenten dar, d. h. das Tagset der TEI wurde hinsichtlich der verfügbaren Elemente und Attribute reduziert und hinsichtlich der Attribut-Werte spezifiziert (Haaf, Geyken & Wiegand 2014/2015; Geyken et al. 2012). Dadurch ist die volle Kompatibilität auch mit anderen TEI-basierten Projekten gewährleistet.

Das DTABf-Annotationsschema (RNG[3]) für historische Drucke (und weitere Dokumentklassen wie Zeitungen und Handschriften, vgl. Haaf & Schulz 2014; Haaf & Thomas 2016 [2017]) bildet zusammen mit einer umfangreichen Dokumentation und einem Schematron-Regelsatz die Grundlage für die XML-Auszeichnung aller Werke im DTA. Mithilfe von Konvertierungstools können aus DTABf-Dokumenten zahlreiche weitere Formate für die Weiterverarbeitung mit linguistischen Werkzeugen, für Suchmaschinenindizes, zur Präsentation der Texte (z. B. Lesefassungen für verschiedene Medien) und zum Export (z. B. in Zitationsumgebungen, Graphdatenbanken oder im CLARIN-Kontext für WebLicht[4]) automatisch erzeugt werden.

Mit dem DTABf als Leitfaden für die Auszeichnung der vielfältigsten Phänomene in historischen Textressourcen ist eine Brücke geschaffen worden, um

2 TEI P5 Guidelines: http://www.tei-c.org/Guidelines/P5/ (letzter Zugriff: 6. 11. 2017).
3 RELAX NG (RNG): http://relaxng.org/ (letzter Zugriff: 6. 11. 2017).
4 WebLicht: https://weblicht.sfs.uni-tuebingen.de/ (letzter Zugriff: 6. 11. 2017).

auch Editionen im DTA nutzbar zu machen. Das DTABf wird von der Deutschen Forschungsgemeinschaft nicht nur für die Textauszeichnung in historischen Sprachkorpora sondern auch als Basisformat für Editionen empfohlen.[5] Die Quelldateien des DTABf-Schemas und der Dokumentation stehen unter einer freien Lizenz zur Nachnutzung zur Verfügung.[6]

2.2 Die kollaborative Qualitätssicherungsumgebung DTAQ

Jedes Werk im DTA bildet eine Einheit aus drei Komponenten: Metadaten, Bilddigitalisate und Textdigitalisat. In der DTA-Infrastruktur werden diese Daten in entsprechenden Datenbanken und Speichersystemen vorgehalten. Diese werden dann auf der Präsentationsebene miteinander verknüpft. Um den verschiedenen Zugangsweisen, Sichten und Darstellungsformaten gerecht zu werden, sind dem Einspielen eines Werks in die DTA-Infrastruktur zunächst (automatisierte) Vorarbeiten vorangestellt:

1. Die Metadaten werden in eine SQL-basierte Datenbank überführt. Dadurch wird später eine gezielte und äußerst flexible Suche über diese Daten möglich. Da der Bestand des DTA auch von anderen Webdiensten regelmäßig abgefragt wird, werden dafür Metadaten in den Formaten Dublin Core,[7] CMDI[8] und EPICUR[9] erstellt und diese über Schnittstellen wie OAI-PMH,[10] BEACON[11] etc. zur Verfügung gestellt.
2. Die Bilddigitalisate werden für die Darstellung im Web in das JPEG-Format konvertiert. Dabei werden verschiedene Auflösungen einer jeden Dokumentseite erstellt: Für Vorschaubilder, die Einzeldarstellung in der Text-Bild-Ansicht und hochauflösend, um Details genauer betrachten zu können.

5 Vgl. die Handreichungen DFG 2015a und DFG 2015b.

6 Zugänglich unter https://github.com/deutschestextarchiv/dtabf (letzter Zugriff: 6. 11. 2107); vgl. auch Haaf (2017).

7 Dublin Core Metadata Initiative: http://dublincore.org/ (letzter Zugriff: 6. 11. 2017).

8 Component MetaData Infrastructure (CMDI): https://www.clarin.eu/content/component-metadata (letzter Zugriff: 6. 11. 2017).

9 Enhancement of Persistent Identifier Services: Comprehensive Method for Unequivocal Resource Identification (EPICUR): http://www.dnb.de/DE/Wir/Projekte/Archiv/epicur.html (letzter Zugriff: 6. 11. 2017).

10 Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): https://www.openarchives.org/pmh/ (letzter Zugriff: 6. 11. 2017).

11 BEACON: https://de.wikipedia.org/wiki/Wikipedia:BEACON (letzter Zugriff: 6. 11. 2017).

3. Das Textdigitalisat wird in einzelne Textseiten (jeweils ein XML-Dokument) zerlegt, da bei größeren Werken mit mehreren hundert Seiten eine Extraktion der entsprechenden Daten zur Laufzeit (z. B. während des Abrufs einer bestimmten Seite in der Text-Bild-Ansicht) nicht performant möglich wäre. Des Weiteren wird eine Reintextfassung (d. h. die reine Transkription mit Zeilenumbrüchen und Absätzen, aber ohne jegliche Annotation) und eine Dokumentfassung für die Indizierung mit der Suchmaschine DDC erstellt. Da die Seitenzerlegung verlustfrei erfolgt, lässt sich aus den Einzelseiten jederzeit wieder das gesamte XML-Dokument erzeugen.

DTAQ ist eine Webanwendung, die weit mehr als nur die Präsentation einzelner Werke ermöglicht. Sie beinhaltet eine Nutzerverwaltung, die mithilfe von Rollen und Rechten mehrere Ebenen von Zugriffen und Annotationsoptionen für verschiedene Nutzergruppen anbietet. Nutzerinnen und Nutzer von DTAQ registrieren sich mit einem personalisierten Account auf der Plattform und können dabei verschiedene Expertisen (Expertise in Literatur- bzw. Sprachgeschichte, Fremdsprachenkenntnisse, Fachkenntnisse in der Transkription mathematischer Formeln u. a.) angeben. Das ermöglicht es, bei Zweifelsfällen oder schwierigen Textstellen gezielt andere Nutzerinnen und Nutzer mithilfe des Ticketsystems anzusprechen und so kollaborativ an den Dokumenten zu arbeiten. Zudem ist so eine Arbeit im Team problemlos möglich, indem bestimmte Arten von Fehlern gezielt einzelnen Nutzer und Nutzerinnen zugewiesen werden können. Die Personalisierung ermöglicht es auch, die eigenen Wünsche in Hinsicht auf die Darstellung in DTAQ für jeden Account zu speichern, unter anderem die optimale Text- und Bildbreite oder die präferierte Textansicht u. v. a.

Als ‚aktives Archiv' ermöglicht es das Deutsche Textarchiv mittels DTAQ, dass Nutzerinnen und Nutzer auch Bearbeitungen an den Textdigitalisaten selbst vornehmen können. Die Ablage der XML-Quellen in einem Versionierungssystem (git[12]) gewährleistet dabei zu jeder Zeit die Transparenz und Nachvollziehbarkeit der Genese eines digitalen Werks. Mithilfe zweier Online-Editoren (im sogenannten WYSIWYG-Modus[13] oder auch direkt im XML-Quelltext) können Nutzerinnen und Nutzer nachträglich Fehler korrigieren sowie zusätzliche Annotationen hinzufügen.

12 git: https://git-scm.com/ (letzter Zugriff: 6. 11. 2017).
13 WYSIWYG: What you see is what you get (man kann also direkt in der Präsentationsform der Texte ändern).

Technisch steht hinter DTAQ ein Webframework (Perl[14], Catalyst[15]), das mithilfe einer PostgreSQL-Datenbank[16] und XML-/XSLT[17]-Tools ein komfortables, effizientes Arbeiten mit dem Dokumentenbestand ermöglicht.

2.3 Erstellen und Kuratieren DTABf-konformer Textressourcen

Das DTA bietet seinen Nutzerinnen und Nutzern eine Vielzahl von Hilfsmitteln, um (historische) Textressourcen zur Integration in die DTA-Infrastruktur von Grund auf neu zu erarbeiten bzw. bestehende Daten zu kuratieren und damit DTABf-konform zu machen. Diese Hilfsmittel decken alle Bereiche der Erstellung bzw. Bearbeitung digitaler Volltexte ab: Verschiedene Werkzeuge bzw. Webservices unterstützen Forscherinnen und Forscher von der Metadatenerfassung über die Transkription und weitere Annotation der Volltextdaten sowie deren Konvertierung in eine lesefreundliche HTML-Ansicht. Das DTA unterstützt somit den gesamten Lebenszyklus digitaler Dokumente von der Erstellung über die Datenkonvertierung und -kuration (*Digital Curation*, vgl. dazu im Kontext des DTA/CLARINs Thomas & Wiegand 2015) bis hin zur Publikation und Analyse und dient damit als Textbearbeitungs- und Analyseplattform für (historische) Textressourcen. Anschließend an die Integration erfolgt die automatisierte computerlinguistische Erschließung der Daten, die somit unmittelbar im Kontext des DTA-Korpus genutzt werden können (siehe Abschnitt 3). Durch diese und weitere Hilfsmittel, zusätzlich unterstützt durch die mit zahlreichen Beispielen versehenen Richtlinien zur Transkription[18] sowie die umfangreiche, ebenfalls mit Beispielen aus den DTA-Korpora illustrierte Dokumentation zum DTABf,[19] werden Nutzerinnen und Nutzer somit in die Lage versetzt, Schritt für Schritt ihre Textressourcen standardkonform aufzubereiten.

Um Primärquellen in DTABf-konformer Weise von Grund auf neu zu erfassen, bietet es sich an, mit der über die Seiten des DTABf bereitgestellten XML-Vorlagendatei[20] zu beginnen und diese sukzessive um die Metadaten und

14 The Perl Programming Language: https://www.perl.org/ (letzter Zugriff: 6. 11. 2017).
15 Catalyst Web Framework: http://www.catalystframework.org/ (letzter Zugriff: 6. 11. 2017).
16 PostgreSQL: https://www.postgresql.org/ (letzter Zugriff: 6. 11. 2017).
17 Extensible Stylesheet Language Transformations (XSLT): https://www.w3.org/TR/xslt (letzter Zugriff: 6. 11. 2017).
18 Siehe http://www.deutschestextarchiv.de/doku/basisformat/transkription (letzter Zugriff: 6. 11. 2017).
19 Siehe http://www.deutschestextarchiv.de/doku/basisformat (letzter Zugriff: 6. 11. 2017).
20 Siehe http://www.deutschestextarchiv.de/files/vorlage_basisformat.xml (letzter Zugriff: 6. 11. 2017).

annotierte Transkription der Quelle zu ergänzen. Die Vorlagendatei ist TEI-basiert und bildet die vorgeschriebene Grundstruktur jedes TEI-Dokuments[21] ab. In der Datei bereits eingebunden sind bereits das DTABf-RNG-Schema, gegen das die Dokumente hinsichtlich der verwendeten Elemente und Attribut-Wert-Paare sowie deren korrekter Strukturierung und Schachtelung validiert werden, und ein entsprechender Schematron-Regelsatz, der weitere formale Festlegungen enthält.[22] Bei handschriftlichen Vorlagen sollte das spezifizierte RNG-Schema zum DTABf für Manuskripte (DTABf-M)[23] eingebunden werden. Als zusätzliche Arbeitserleichterung steht ein Webformular zur Erstellung DTABf-konformer TEI-Header zur Verfügung, über das die relevanten Angaben komfortabel erfasst werden können und aus dem heraus schließlich per Knopfdruck ein vollständiger, in sich valider TEI-Header erzeugt werden kann.[24]

Auch für die auf die Erstellung des XML-Dokuments folgenden Arbeitsschritte bietet das DTA hilfreiche Tools und Anwendungen an: Beispielsweise können die im DTA entwickelten XSLT-Stylesheets, mit deren Hilfe die XML-Datei in ein HTML-Format umgewandelt wird, genutzt werden, um eine lesefreundliche, an die spätere Darstellung auf der DTA-Seite angelehnte ‚Vorschauansicht' des DTABf-Dokuments im Browser zu generieren. Andere kleine Tools dienen beispielsweise dazu, die fortlaufende Nummerierung der aufeinanderfolgenden `<pb/>`-Elemente (d. h. die fortlaufende Nummerierung der Digitalisate der Vorlage) in der Datei (wieder-)herzustellen – eine ansonsten mühsame Handarbeit – oder die im Dokument enthaltenen Sonderzeichen[25] automatisch in die entsprechenden numerischen Entitäten für Zeichenverweise in der Form `&#xNNNN;` zu überführen.

21 Vgl. http://www.deutschestextarchiv.de/doku/basisformat/grundstrukturDokument (letzter Zugriff: 6. 11. 2017).

22 Vgl. allgemein Regular Language Description for XML New Generation (RELAX NG) und ISO Schematron, unter: http://relaxng.org/; http://schematron.com/ (letzter Zugriff: 6. 11. 2017). Speziell zu deren Verwendung und Handhabung innerhalb des DTABf die Dokumentation unter http://www.deutschestextarchiv.de/doku/basisformat/benutzungDTABfSchema (letzter Zugriff: 6. 11. 2017) sowie Haaf (2017).

23 Vgl. zum DTA-Basisformat für Manuskripte (DTABf-M) die Dokumentation unter http://www.deutschestextarchiv.de/doku/basisformat/manuskript (letzter Zugriff: 6. 11. 2017) sowie Haaf & Thomas (2016 [2017]).

24 Verfügbar unter http://www.deutschestextarchiv.de/dtae/submit/clarin (letzter Zugriff: 6. 11. 2017).

25 D. h. diejenigen, die außerhalb des Bereichs der am häufigsten verwendeten Unicode-Zeichen (kleinergleich U+00FF) liegen und daher oft zu Darstellungs-, Verarbeitungs- oder Dekodierungsfehlern führen.

In ähnlicher Weise wie für die im bisherigen Teil dieses Abschnitts beschriebene Neuerstellung DTABf-konformer Ressourcen können die vorgestellten Hilfsmittel – allen voran das RNG-Schema und der Schematron-Regelsatz, ggf. aber auch die hier vorgestellten Skripte, Tools und Webservices – zur Kuratierung externer Ressourcen genutzt werden. Dabei werden externe Textressourcen in das DTABf konvertiert und anschließend via DTAQ in die Korpusinfrastruktur integriert. Beispielsweise wurden im Rahmen eines von CLARIN-D geförderten Kurationsprojekts historische Textressourcen des 15.–19. Jahrhunderts aus verschiedenen Quellen in das DTA integriert.[26] (Konkrete Beispiele für integrierte externe Ressourcen finden sich in Abschnitt 4.2.)

Ein weiteres, im Rahmen von DTAQ bereitgestelltes Hilfsmittel dient dazu, die auf dem beschriebenen Weg erstellten, DTABf-konform annotierten Dokumente hinsichtlich der historischen Schreibweisen zu ‚normalisieren'.[27] Dieser Service spricht den CAB-Webservice[28] an, um für die historischen Schreibweisen im Text ein modernes Äquivalent zu ermitteln. Die Anreicherung des Textes mit den modernisierten Formen wird gemäß editorischen Konventionen mit Hilfe der TEI-Elemente `<choice>`, `<orig>` und `<reg>` dokumentiert. Das Element `<reg>` wird dabei zusätzlich mit dem Attribut-Wert-Paar `@resp="#cab"` als automatisierter, d. h. eben vom Webservice CAB verantworteter Eingriff gekennzeichnet. Beispielsweise wird die Transkription der Phrase „EJne feſte Burgk iſt vnſer GOtt/ ꝛc.“[29] vollautomatisch annotiert als:

```
<choice><orig>EJne</orig>
    <reg resp="#cab">Eine</reg></choice>
<choice><orig>feſte</orig>
    <reg resp="#cab">feste</reg></choice>
<choice><orig>Burgk</orig>
    <reg resp="#cab">Burg</reg></choice>
<choice><orig>iſt</orig>
```

26 Vgl. zu diesem Projekt http://deutschestextarchiv.de/doku/clarin_kupro_publikationen (letzter Zugriff: 6. 11. 2017) bzw. Thomas & Wiegand (2015).

27 Dieser Webservice sowie alle im vorhergehenden Absatz erwähnten und weitere DTAQ-Tools sind verfügbar unter http://www.deutschestextarchiv.de/dtaq/tool (letzter Zugriff: 6. 11. 2017).

28 Vgl. DTA::CAB Web Service v1.82, http://www.deutschestextarchiv.de/cab/ (letzter Zugriff: 6. 11. 2017).

29 Vgl. zu diesem leicht modifizierten Textbeispiel [N. N.]: Jubilaeum Typographorum Lipsiensium Oder Zweyhundert-Jähriges Buchdrucker JubelFest. [Leipzig], 1640, S. [19]. In: Deutsches Textarchiv, http://www.deutschestextarchiv.de/oa_jubilaeum_1640/27 (letzter Zugriff: 6. 11. 2017).

```
    <reg resp="#cab">ist</reg></choice>
<choice><orig>vnſer</orig>
    <reg resp="#cab">unser</reg></choice>
<choice><orig>GOtt</orig>
    <reg resp="#cab">Gott</reg></choice>/
<choice><orig>ꝛc.</orig>
    <reg resp="#cab">etc.</reg></choice>
```

Auf diese Weise können auf denkbar einfache Art und unabhängig von bzw. noch vor der Integration der Texte in die DTA-Infrastruktur schreibweisennormierte Fassungen der vorlagengetreuen Transkriptionen historischer Textzeugen erstellt werden, die dann wiederum zur Bearbeitung mit externen, für gegenwartssprachliche Texte optimierten Anwendungen – beispielsweise zur Eigennamenerkennung oder Topic Modeling – genutzt werden können.

2.4 Arbeiten in DTAQ

Die Veröffentlichung der Ressourcen erfolgt zunächst auf der passwortgeschützten DTAQ-Plattform. Hier findet die summative Qualitätssicherung[30] statt, d. h. die Text-Bild-Zuordnung, die Transkription und Annotation usw. können webbasiert und kollaborativ geprüft werden. Für jedes Dokument wird eine eigene Startseite erstellt, auf der die bibliographischen Metadaten und weitere relevante Informationen zur Quelle, einschließlich der zugrunde gelegten Text- und Bildvorlagen, der Lizenz für die Nachnutzung, der Korpuszugehörigkeit des Dokuments, der Bearbeiter der digitalen Edition sowie die dabei ggf. verbliebenen Abweichungen der Transkription bzw. der Annotation von den DTA-Vorgaben zusammengefasst sind. Zu beiden Bereichen, d. h. den allgemeinen Informationen und den bibliographischen Metadaten, können über einen Browserdialog Anmerkungen und Berichtigungen hinzugefügt werden. Unter *Ansichten* stehen in der Korrekturumgebung DTAQ bzw., sobald das Dokument freigeschaltet ist, auf der DTA-Webseite neben dem Zugang zur Text-Bild-Ansicht unter anderem verschiedene, automatisch aus dem TEI-XML generierte Download-Formate sowie eine Reihe analytischer Zugänge bereit. Mit der linguistischen Suchmaschine DDC und der grep-Suche werden parallel zwei Volltextsuchen angeboten.

30 Vgl. für eine ausführliche Darstellung der formativen und summativen Qualitätssicherung im DTA Geyken et al. 2012.

Ein automatisch aus der `<div>`-Strukturierung des Dokuments erzeugtes *Inhaltsverzeichnis* erlaubt das gezielte Ansteuern bestimmter Textpassagen. Durch Klick auf einen Inhaltsabschnitt, auf die Ansicht *Korrekturumgebung* oder auf das Faksimile des Titelblatts gelangt man von dieser Startseite aus zur Text-Bild-Ansicht in DTAQ. Neben dem Faksimile der Vorlage in der ersten Spalte der dreiteiligen Ansicht wird in der mittleren Spalte die Transkription des Textes der entsprechenden Seite angezeigt. Für die Textansicht sind verschiedene Optionen wählbar: die XML-Ansicht, eine reine Textansicht, die CAB-Ansicht mit der automatisch durchgeführten orthographischen Normalisierung der zugrundeliegenden Transkription, eine erweiterte Ansicht der Part-of-Speech-Analyse sowie der Lemmatisierung.

Zur Überprüfung des Texts eignet sich die standardmäßig angezeigte HTML-Fassung am besten.[31] In dieser lesefreundlichen, per XSLT aus dem XML erstellten Textansicht lassen sich nun die bei dem jeweiligen Korrekturgang gefundenen Transkriptions-, Auszeichnungs- oder Druckfehler mittels sogenannter Tickets melden. Die gemeldeten Tickets enthalten neben dem Fehlertyp, der fehlerhaften Textstelle, einem Vorschlag zur Behebung des Fehlers und ggf. weiteren Kommentaren zugleich auch die Angaben zum Zeitpunkt der Erfassung und zum DTAQ-Account, durch den die Meldung angelegt wurde. Beobachtungen, die sich nicht nur auf die jeweilige Seite, sondern das gesamte Dokument beziehen, können entsprechend gekennzeichnet werden. Darüber hinaus ist es möglich, dem Ticket eine mehr oder weniger hohe Priorität zuzuordnen und dieses einer bestimmten Bearbeiterin bzw. einem bestimmten Bearbeiter zuzuweisen. Anschließend wird der Korrekturstatus der Seite gekennzeichnet, d. h. ob die Überprüfung auf der Ebene des Textes, eines punktuellen Vergleichs von Transkription und Faksimile und/oder auf der Ebene der XML-Annotation durchgeführt wurde. Die gemeldeten Fehler werden so anhand der Beschreibungen in den Tickets vom DTA-Team geprüft und entsprechend direkt in den XML-Dokumenten behoben.

Eine andere Möglichkeit, die auch durch DTA-externe Nutzerinnen und Nutzer zur sukzessiven Korrektur bzw. Kuratierung der Ressourcen direkt in der webbasierten DTAQ-Oberfläche genutzt werden kann, sind die beiden integrierten Online-Editoren, die optional bereitgestellt werden. Der einfache WYSIWYG-Editor bietet die Möglichkeit, innerhalb der HTML-Ansicht Änderungen auf der Textoberfläche, insbesondere also bei Transkriptionsfehlern, vorzunehmen. Die vorgenommenen Änderungen werden im Zuge der Speicherung automatisch dokumentiert und versioniert. Der parallel dazu angebotene XML-

31 Vgl. zu diesem Abschnitt abermals Geyken et al. (2012) sowie Haaf & Thomas (2016: insbes. 227 f.).

Editor erlaubt Änderungen, die (auch) das strukturelle und typographische Markup der Dokumente betreffen; hier können beispielsweise manuelle Normalisierungen vorgenommen oder Druckfehler in dokumentierter Weise mittels der TEI-Elemente `<choice>`, `<sic>` und `<corr>` behoben werden:

Abb. 10.2: XML-Editor in DTAQ mit geöffnetem Browserdialog (Wrap-Tag-Funktion), http://www.deutschestextarchiv.de/dtaq/book/view/humboldt_manati_1838?p=9&view=xmleditor (letzter Zugriff: 6. 11. 2017)

Neben der farblich differenzierten Syntaxhervorhebung und der Wrap-Tag-Funktion, die Nutzerinnen und Nutzer von etablierten Programmen wie dem oXygen XML Editor gewohnt sind, bietet die rechte Spalte eine Anzahl häufig genutzter Sonderzeichen und Tags zur Auswahl, die die Bearbeitung der Dokumente direkt im XML-Modus erleichtern. Vorgenommene Änderungen auf der XML-Ebene werden zunächst auf ihre Wohlgeformtheit und auf ihre Validität hin gegen das DTABf-Schema geprüft, um sicherzustellen, dass keine in dieser Hinsicht ungültigen Änderungen in den XML-Dokumenten vorgenommen werden; anschließend werden die formal korrekten Änderungen im XML-Dokument gespeichert und ebenfalls versioniert.

Durch die beiden webbasierten, direkt in DTAQ eingebundenen Editoren lassen sich nun nicht nur die bereitgestellten Quellen sehr viel komfortabler und mit Unterstützung auch DTA-externer Nutzerinnen und Nutzer korrigieren. Sie ermöglichen darüber hinaus auch eine effiziente und fortlaufende Verfeinerung bzw. Vertiefung der Annotation insgesamt. So wurden beispielsweise im

Kooperationsprojekt *Hidden Kosmos*[32] zunächst alle Manuskripte des Korpus DTABf-konform transkribiert, hinsichtlich der wesentlichen Struktur- und graphematischen Merkmale ausgezeichnet und in DTAQ publiziert. Anschließend erfolgte nicht nur die Qualitätssicherung und Korrektur von Transkriptions- und Auszeichnungsfehlern direkt in DTAQ, sondern wurden innerhalb des etwa 3.500 Seiten umfassenden Korpus auch sämtliche vorkommenden Personennamen explizit ausgezeichnet. Die insgesamt mehr als 8.000 Personennamen wurden mit einem `<persName>`-Tag versehen und, soweit möglich, mittels des `@ref`-Attributs mit einem eineindeutigen Identifizierer aus einem Normdatensatz[33] versehen. Das so erzeugte, umfassende Personenregister[34] konnte dank des webbasierten XML-Editors ortsunabhängig vom Projektteam an der Humboldt-Universität erstellt werden. In ähnlicher Weise wurden innerhalb des DFG-Projekts *AEDit Frühe Neuzeit*[35] etwa 340 Leichenpredigten mit mehr als 16.000 Seiten im Team der Kooperationspartner BBAW, Herzog August Bibliothek Wolfenbüttel (HAB) und der Forschungsstelle für Personalschriften an der Philipps-Universität Marburg von den verschiedenen Standorten aus kollaborativ online bearbeitet.

3 Analysewerkzeuge im DTA

3.1 Linguistische Datenanalyse im DTA

Für bestimmte Anwendungsbereiche der Auswertung der Korpora im DTA werden bereits über die DTA-Plattform eigene Werkzeuge angeboten. Auf diese Weise können die DTA-Daten hinsichtlich verschiedener Phänomentypen analysiert werden, ohne dass sie heruntergeladen und mit externen Tools ver-

32 Vgl. Humboldt-Universität zu Berlin: *Hidden Kosmos – Reconstructing Alexander von Humboldt's "Kosmos-Lectures"*, http://www.culture.hu-berlin.de/hidden-kosmos (letzter Zugriff: 6. 11. 2017).

33 Genutzt wurde vorzugsweise die Gemeinsame Normdatei (GND), vgl. http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (letzter Zugriff: 6. 11. 2017); sofern keine Normdaten zu den betreffenden Personen in der GND verzeichnet waren, wurden alternativ die Datensätze aus dem Virtual International Authority File, https://viaf.org/ (letzter Zugriff: 6. 11. 2017) oder Wikidata, https://www.wikidata.org/wiki/Wikidata:Hauptseite (letzter Zugriff: 6. 11. 2017), verwendet.

34 Verfügbar unter http://www.deutschestextarchiv.de/kosmos/person (letzter Zugriff: 6. 11. 2017).

35 Vgl. dazu http://www.deutschestextarchiv.de/doku/textquellen#aedit (letzter Zugriff: 6. 11. 2017).

knüpft werden müssen. Das Potential der linguistischen Analyse auf der DTA-Plattform wird im Folgenden umrissen.

Sämtliche Texte des DTA durchlaufen vollautomatisch eine Reihe linguistischer Verarbeitungsschritte, welche durch die Software CAB geleistet werden. CAB umfasst die Satzsegmentierung, Tokenisierung, Lemmatisierung, Modernisierung historischer Schreibweisen sowie das Part-of-Speech-Tagging gemäß STTS (Jurish 2012; Jurish & Würzner 2013).

Die Ergebnisse der linguistischen Analyse können mithilfe der Suchanfragesprache DDC[36] in die Korpusanalyse eingebunden werden. DDC ermöglicht die Formulierung komplexer Suchanfragen (Jurish, Thomas & Wiegand 2014).[37] Insbesondere können damit linguistische Annotationen auf Wortposition (Lemmatisierung, POS-Tagging) mit Phrasensuche, Boole'schen Operatoren und regulären Ausdrücken verknüpft werden.

Um die Analyseergebnisse für ein Token einzusehen sowie um zu überprüfen, welches Lemma, Wortart bzw. welche orthographischen Varianten für dieses Token automatisch ermittelt wurden, kann die Benutzeroberfläche des CAB-Webservice[38] konsultiert werden. Hier ist z. B. einsehbar, dass das Token „Frewde" dem Lemma „Freude" sowie dem POS-Tag „NN" (Appellativum) zugeordnet wird und es auf eine Vielzahl von möglichen orthographischen Varianten abgebildet wird, von „Frewdt" über „fräud" und „Freüd" bis hin zu „fröude", „fröide" oder „vröide".[39] Vor allem aber kann der CAB-Webservice genutzt werden, um eigene Texte mit der CAB-Software zu analysieren. Für einen schnellen Einblick in die orthographische Normierung und das POS-Tagging im größeren Textzusammenhang steht die CAB- bzw. POS-Ansicht zu jeder Buchseite in DTAQ zur Verfügung.

---

36 Dialing DWDS Concordancer (DDC), http://www.deutschestextarchiv.de/doku/software#ddc; vgl. Jurish, Thomas & Wiegand 2014. Zu den vielfältigen Möglichkeiten der Volltextsuche im DTA mit der DDC-Suchmaschine siehe auch die Hilfeseite http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe (letzter Zugriff: 6. 11. 2017).

37 Für die DDC-basierte Recherche in allen Korpora des Deutschen Textarchivs steht die Suchmaske unter http://kaskade.dwds.de/dstar/dta (letzter Zugriff: 6. 11. 2017) zur Verfügung. Hier lässt sich einschränken, ob im Gesamtkorpus oder in ausgewählten Teilkorpora recherchiert werden soll (sog. Flags). Die Korpora umfassen auch alle Texte in DTAQ. Zur DDC-Syntax vgl. als Einstieg http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe (letzter Zugriff: 6. 11. 2017) sowie weiterführend http://odo.dwds.de/~jurish/software/ddc/querydoc.html (letzter Zugriff: 6. 11. 2017).

38 Vgl. Anm. 27; Einstieg und Links zur Dokumentation: http://odo.dwds.de/~moocow/software/DTA-CAB/doc/html/DTA.CAB.WebServiceHowto.html (letzter Zugriff: 6. 11. 2017).

39 Vgl. http://www.deutschestextarchiv.de/demo/cab/query?a=expand&fmt=text&raw=1&q=Frewde (letzter Zugriff: 6. 11. 2017).

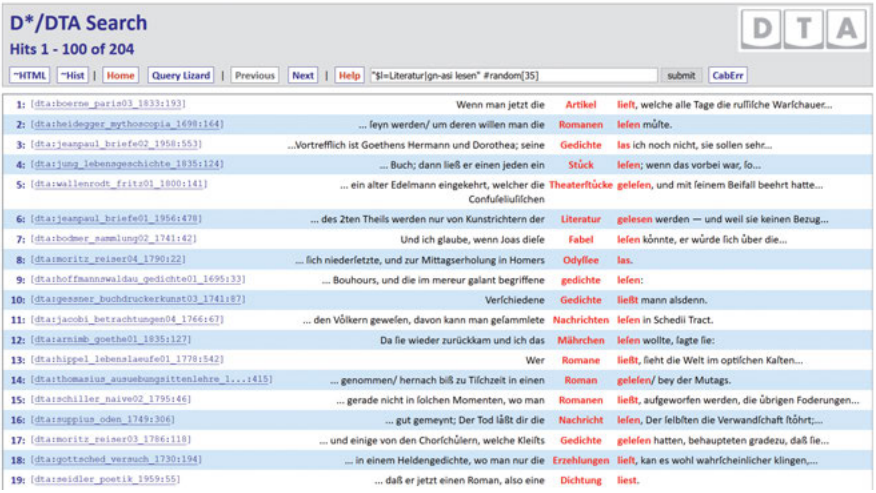

Abb. 10.3: DDC-Abfrage: `"$l=Literatur|gn-asi lesen"`.

Alle Texte der DTA-Korpora wurden mit einer Thesaurus-Funktion basierend auf dem Wortnetz GermaNet versehen. Über die Verknüpfung der Lemmata mit SynSets ist daher nun eine semantische Recherche möglich. Konkret können für ein Lemma dessen Hyperonyme, Hyponyme oder Synonyme im Korpus recherchiert werden (Abb. 10.3).

Valide Expansionen sind dabei `"gn-asi"` (sämtliche Hyponyme zum gegebenen Lemma), `"gn-isa"` (entsprechend die Hyperonyme), `"gn-syn"` (entsprechend die Synonyme). Die Suche kann auch auf die Hypero- und Hyponyme bestimmter Ordnungen (`"gn-asi1"`, `"gn-asi2"`) eingeschränkt werden. Neben GermaNet ist darüber hinaus auch der OpenThesaurus[40] an die DTA-Korpora angebunden, sodass äquivalente Recherchen auf dem Datenmaterial dieses Thesaurus möglich sind.[41]

3.2 Visualisierung des zeitlichen Verlaufs mit Verlaufskurven

Über die reine Korpusrecherche hinaus sind verschiedene lexikometrische Analysen über die DTA-Plattform möglich. So lässt sich etwa die relative Verteilung von (komplexen) Ausdrücken in den DTA-Korpora mit Hilfe einer Wort-

40 Vgl. https://www.openthesaurus.de (letzter Zugriff: 6. 11. 2017).

41 Für eine Dokumentation zu den Thesaurus-Funktionalitäten im DTA vgl. http://odo.dwds.de/~moocow/software/ddc/querydoc.html#dta_expand_gn (letzter Zugriff: 6. 11. 2017) bzw. http://odo.dwds.de/~moocow/software/ddc/querydoc.html#dta_expand_ot (letzter Zugriff: 6. 11. 2017).

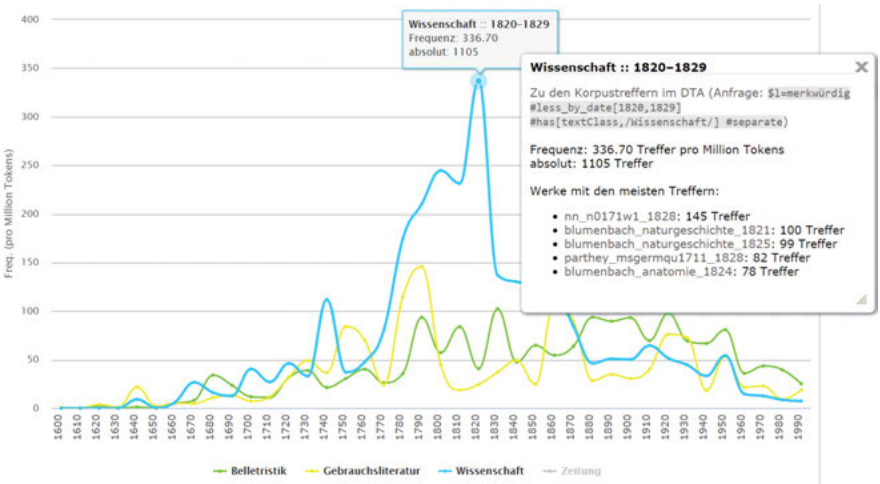

Abb. 10.4: Wortverlaufskurve für DDC-Abfrage: `$l=merkwürdig`.

verlaufskurve darstellen, welche auf normierten Frequenzberechnungen und Glättungsverfahren beruht (Geyken et al. 2015).[42] Für das Lemma „merkwürdig“ wird so z. B. auf einen Blick deutlich, dass es in der Wissenschaftssprache des 19. Jahrhunderts signifikant häufig verwendet wurde (Abb. 10.4). Neben der Standardansicht kann auch eine Ansicht mit den Rohfrequenzen gewählt werden. In dieser lassen sich für jedes Jahr und jede Textsorte die absoluten Frequenzen nachvollziehen. Schließlich ermöglicht die erweiterte Ansicht die genaue Einstellung aller Parameter, wie beispielsweise der Zeitintervalle, der Glättungsumgebung oder des Konfidenzintervalls zur Bestimmung von Ausreißern. Eine weitere Besonderheit der Wortverlaufskurve besteht darin, dass alle Ergebnisse an die Korpora zurückgebunden werden. Durch Klick auf jeden Messpunkt gelangt man zu den Korpus-Konkordanzen. Im Falle der Abbildung 10.4 kann man somit beispielsweise durch Klick auf das Maximum im Zeitintervall 1820–1829 erfahren, dass unter anderem die Nachschriften der Kosmos-Vorlesungen Alexander von Humboldts sowie Werke Johann Friedrich Blumenbachs hinter diesen hohen Trefferzahlen stehen.

42 Die Wortverlaufskurve ist zugänglich unter http://www.deutschestextarchiv.de/search/plot/ (letzter Zugriff: 6. 11. 2017).

3.3 Kollokationsanalyse mit DiaCollo

Die weiterführende Untersuchung der Entwicklung von Wörtern in ihren lexikalischen Kontexten im zeitlichen Verlauf ist dann mit dem Werkzeug DiaCollo möglich (Jurish 2015; Jurish, Geyken & Werneke 2016, Lemnitzer, Jurish & Burkhardt 2016). DiaCollo ermittelt die statistisch signifikanten Kollokationen zu einem gegebenen Ausdruck in verschiedenen Zeitschnitten und visualisiert ihre Signifikanz mit einem Farbschema. Die Analyse kann dabei individuell parametrisiert werden. Zum Beispiel ist es möglich, die Größe der Zeitschnitte anzupassen, die Kollokate auf bestimmte POS-Tags zu beschränken oder die Menge der einbezogenen stärksten Kollokate (k-best-Wert) zu variieren. Verschiedene Visualisierungen stehen zur Verfügung, wie beispielsweise eine *bubble*-Ansicht, eine Ansicht als Schlagwortwolke oder die von Google entwickelte *gmotion*-Ansicht.

Nutzt man z. B. DiaCollo, um die Nomen zu ermitteln, welche typische Kollokationen des Tokens „merkwürdig“ sind, so zeigt sich zunächst, dass ganz generell die Menge solcher typischen Nomen-Kollokationen zunimmt (d. h. dass sich die Kontexte, in welchen dieses Adjektiv typischerweise verwendet werden konnte, vervielfältigen). Des Weiteren wird deutlich, dass zwischen 1750 und 1790 die „merkwürdige Begebenheit“ eine recht typische Wortverbindung war, während ab den 1820er Jahren bemerkenswert häufig von der „merkwürdigen Erscheinung“ die Rede ist (Abb. 10.5). Ebenso wie bei den in Ab-

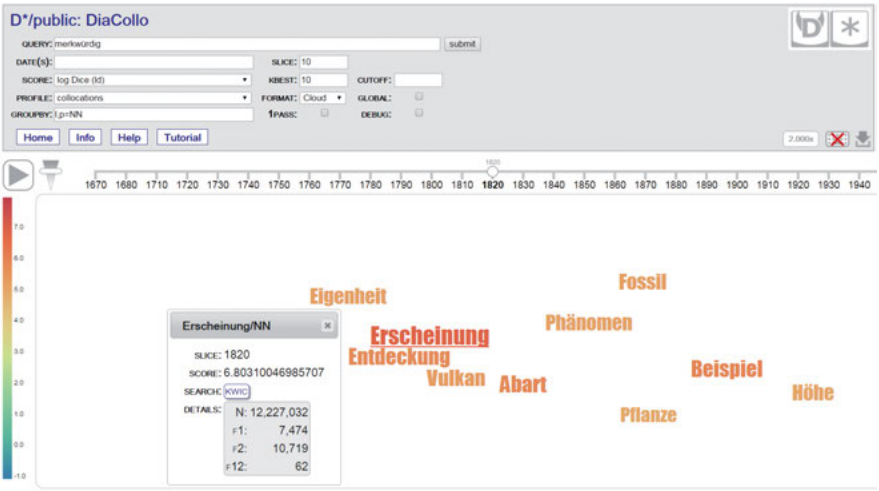

Abb. 10.5: DiaCollo: typische Nomen-Kollokationen (GROUPBY: l,p = NN) zu „merkwürdig“ im Zeitschnitt 1820.

schnitt 3.2 beschriebenen Wortverlaufskurven ermöglicht auch DiaCollo die Rückbindung an die einzelnen Korpuskonkordanzen durch Klick auf die Kollokate. In Abbildung 10.5 würde beispielsweise der Klick auf das Kollokat „Erscheinung“ zu einer Trefferliste mit den Konkordanzen führen.

3.4 Quantitative Einzeltextanalyse mit Voyant Tools

Schließlich werden die einzelnen Dokumente an die Voyant Tools[43] angebunden, die für das jeweilige DTA-Werk über einen entsprechenden Unterpunkt auf der Buchstartseite erreicht werden können. Die XML-Volltexte aus dem DTA werden eigens zu diesem Zweck und ohne weiteren nutzerseitigen Aufwand präprozessiert, um eine nahtlose Verwendung und optimale Analyseergebnisse zu gewährleisten. Zur Analyse mit Voyant stellt das DTA drei spezielle XML-Fassungen zur Verfügung:

1. Eine *zeichennormierte Fassung* (unicruftxml): Diese XML-Fassung bietet den Text in transliterierter Orthographie, d. h. in einer Fassung, in der alle Zeichen, die außerhalb der Latin-1-Kodierung (ISO/IEC 8859-1) liegen, durch Zeichen innerhalb von Latin-1 approximiert werden. Damit sind Probleme bei der Voyant-seitigen Behandlung von Zeichen wie dem ‚langen s‘ (ſ, U+017F) oder dem ‚hochgestellten e‘ (U+0364) zur Kennzeichnung von Umlauten ausgeschlossen. Abgesehen davon bleiben die Graphie der Vorlage und auch die Silbentrennung am Seiten- und Zeilenende erhalten.
2. Eine *hinsichtlich der Schreibweisen normierte Fassung* (normxml): Diese XML-Fassung bietet den Text ebenfalls Latin-1-approximiert (siehe 1.) und zusätzlich in normalisierter Orthographie basierend auf der Verarbeitung mit CAB. In diesem Zuge wird auch die Silbentrennung am Seiten- und Zeilenumbruch aufgelöst.
3. Eine *lemmatisierte Fassung* (lemmaxml): Diese XML-Fassung bietet den Text in lemmatisierter Forma, wobei für die Lemmata ebenfalls mit normierten Zeichen (nach 1.) und modernisierter Orthographie (nach 2.) wiedergegeben werden.

Über die Voyant Tools sind nun verschiedene Frequenzanalysen auf Dokumentebene visualisierbar, so etwa Term- und Phrasenfrequenzen sowie Frequenzentwicklungen für Terme (Trends) im Verlauf des Dokuments (Sinclair & Rockwell 2017). So können Terme hinsichtlich ihrer Bedeutung für das jeweilige Werk analysiert werden (Vgl. z. B. Bird, Menzies & Zimmermann 2015: 51).

43 Vgl. https://voyant-tools.org/ (letzter Zugriff: 14. 5. 2018).

Abb. 10.6: Voyant-Analyse von Keller, Gottfried: Der grüne Heinrich. Bd. 1. Braunschweig, 1854, basierend auf der normalisierten Textfassung, unter: http://www.deutschestextarchiv.de/book/download_normxml/keller_heinrich01_1854 (letzter Zugriff: 6. 11. 2017).

4 Interdisziplinäre Vernetzung des DTA

4.1 Verwertung von DTA-Daten in externen wissenschaftlichen Kontexten

Die vorgestellten Analysewerkzeuge der DTA-Plattform ermöglichen bereits die Auswertung der DTA-Korpora in vielerlei Hinsicht. Zudem ist die weitere Nachnutzung der Daten mithilfe externer Werkzeuge, die z. B. über die CLARIN-Infrastruktur bereitgestellt werden, möglich und selbstverständlich auch erwünscht. Wesentliche Voraussetzung dafür ist zum einen die Interoperabilität der Daten untereinander, sodass diese vollautomatisch und jederzeit mit geringem Aufwand in die jeweiligen Eingabeformate konvertiert werden können. Zum anderen ist diese Form der Nachnutzung besonders gut möglich, wenn sowohl bei den Daten als auch bei den Eingabeformaten der Werkzeuge standardisierte Formate zur Anwendung kommen, die dokumentiert und leicht zugänglich sind. Diese beiden Voraussetzungen werden für die Korpusdaten des DTA durch das DTA-Basisformat erfüllt, das eine einheitliche und eindeutige Richtlinie für die Textauszeichnung darstellt und dabei auf den weit verbreiteten TEI-Richtlinien basiert.

Für die freie Nachnutzung in externen Kontexten werden die DTA-Texte einzeln und als Korpus zum Download zur Verfügung gestellt.[44] Da sämtliche Daten einheitlich entsprechend dem DTA-Basisformat kodiert sind, können sie mit geringem Aufwand in andere Formate konvertiert werden. Über die DTA-Plattform werden die Texte außerdem bereits im TCF (*Text Corpus Format*; Heid et al. 2010) bereitgestellt, dem Eingangsformat für die CLARIN-Services (Eckart 2012). Die bereitgestellten TCF-Dateien können zum Beispiel in die WebLicht-Plattform (Hinrichs, Hinrichs & Zastrow 2010)[45] hineingeladen und dort mit den Tools verschiedener CLARIN-Zentren weiter analysiert werden. So stehen über WebLicht zum Beispiel weitere Tools zur linguistischen Vorverarbeitung (Tokenisierung, Lemmatisierung, POS-Tagging etc.), verschiedene Syntaxparser (z. B. Stanford Parser) sowie Tools zur morphologischen Analyse und zur Eigennamenerkennung zur Verfügung. Bei der Nutzung von WebLicht können verschiedene Werkzeuge in Folge zur Anwendung kommen, wobei die Zusammensetzung der jeweiligen Werkzeugkette individuell bestimmt werden kann. Um genuine TEI-XML-Dateien in WebLicht weiterverarbeiten zu können, stellt die BBAW einen Converter als Webservice über WebLicht bereit, der die Konvertierung aus TEI nach TCF und die Rückkonvertierung nach TEI im Anschluss an die WebLicht-Analyse leistet.[46]

Des Weiteren sind die DTA-Korpora an die Föderierte Volltextsuche (*Federated Content Search*)[47] von CLARIN angebunden und somit gemeinsam mit anderen CLARIN-Korpora direkt recherchierbar. Die Verbindung der DTA-Korpusdaten mit anderen Korpora im CLARIN-Verbund wird außerdem über die CLARIN-übergreifende Metadatenplattform VLO (*Virtual Language Observatory*)[48] erreicht, welche durch die Bereitstellung von CMDI-Metadatensätzen zum Harvesting bedient wird (Wittenburg & Uytvanck 2012).

Auch in anderen Umgebungen außerhalb CLARINs ist die Nachnutzung der DTA-Daten möglich und erprobt. Metadaten werden in den genannten Formaten und zusätzlich im Dublin-Core-Format über die OAI-PMH-Schnittstelle des DTA bereitgestellt und beispielsweise von der Europeana[49] und der Biele-

44 Einzelne Werke können über die jeweilige Buchstartseite in verschiedenen Formaten heruntergeladen werden; der Download des Korpus ist unter http://www.deutschestextarchiv.de/download (letzter Zugriff: 6. 11. 2017) möglich.

45 Zugänglich unter: https://weblicht.sfs.uni-tuebingen.de/ (letzter Zugriff: 6. 11. 2017).

46 Außerhalb WebLicht zugänglich unter: http://kaskade.dwds.de/tei-tcf/ (letzter Zugriff: 6. 11. 2017).

47 Zugänglich unter: https://www.clarin.eu/contentsearch (letzter Zugriff: 6. 11. 2017).

48 Zugänglich unter: http://vlo.clarin.eu (letzter Zugriff: 6. 11. 2017).

49 Vgl. http://www.europeana.eu (letzter Zugriff: 6. 11. 2017).

feld Academic Search Engine (BASE)[50] abgefragt. Darüber hinaus harvesten einige Bibliotheken, welche die physischen Quellen der Werke im DTA vorhalten, Informationen zu den entsprechenden Werken im DTA und verlinken diese über ihren OPAC.[51]

Die DTA-Korpusdaten wurden außerdem bereits in andere Korpustools eingespeist. So wurde die TCF-Fassung des DTA-Korpus für ein entsprechendes Korpus-Add-on des Werkzeugs *Corpus Explorer* nachgenutzt, wobei die Informationen zu Tokens, Lemmata, POS und Orthographie ausgewertet wurden.[52] Auch der *Corpus Explorer* bietet darauf aufbauend Syntax Parsing, Kookkurenzanalysen, Verlaufskurven u. a. an. Die Möglichkeit, mit CAB analysierte Daten in das Korpuswerkzeug TXM zu integrieren und dort weiter zu verarbeiten, wurde für Karl Philipp Moritz' Werk *Anton Reiser* prototypisch realisiert.[53] Eine aktuelle Entwicklung bildet die Weiterentwicklung des Tools JCore, eines Werkzeugs zur Nutzung von NLP-Prozessketten in UIMA[54], ursprünglich mit Fokus auf englischsprachiger biomedizinischer Wissenschaftsliteratur, für die deutsche Sprache. In diesem Zusammenhang entstand der *DTA Collection Reader*[55] (Hahn et al. 2016; Hellrich, Matthies & Hahn 2017).

Das DTA-Kernkorpus wurde zudem in das Leipziger Werkzeug CTS (*Canonical Text Service Protocol*) integriert, wo es mit feingranularen persistenten IDs für die Zitation versehen wird und mit darauf aufbauenden Software-Werkzeugen (z. B. zur Alignierung von Textstellen unterschiedlicher Korpora) weiterverarbeitet werden kann (Tiepmar et al. 2016; Tiepmar et al. 2017). Ferner wurde die TCF-Fassung des DTA-Korpus experimentell in eine Graph-Datenbank integriert (Kuczera 2017). Weiterhin bilden 633 Werke der

50 Vgl. https://www.base-search.net/ (letzter Zugriff: 6. 11. 2017).

51 So etwa die Staatsbibliothek zu Berlin (SBB-PK, http://staatsbibliothek-berlin.de/ [letzter Zugriff: 6. 11. 2017]) oder die Staats- und Universitätsbibliothek Göttingen (https://www.sub.uni-goettingen.de [letzter Zugriff: 6. 11. 2017]); vgl. z. B. die entsprechende Verlinkung Kants *Critik der reinen Vernunft* (1781, http://www.deutschestextarchiv.de/kant_rvernunft_1781), abgerufen am 31. 3. 2017 im OPAC der SBB-PK: http://stabikat.sbb.spk-berlin.de/DB=1/XMLPRS=N/PPN?PPN=83453522X.

52 Vgl. http://notes.jan-oliver-ruediger.de/korpora/; http://notes.jan-oliver-ruediger.de/dta-kernkorpus-als-korpus-addon-verfuegbar/ (letzter Zugriff: 6. 11. 2017).

53 Vgl. https://groupes.renater.fr/wiki/txm-users/public/umr_ihrim_moritz (letzter Zugriff: 6. 11. 2017).

54 Unstructured Information Management Architecture; http://uima.apache.org/ (letzter Zugriff: 6. 11. 2017).

55 Zugänglich unter: https://github.com/JULIELab/jcore-base/tree/master/jcore-dta-reader (letzter Zugriff: 6. 11. 2017).

literarischen Moderne aus den DTA-Korpora einen Teil des Korpus KOLIMO (Herrmann & Lauer 2017).[56]

Für die Editionswissenschaften wurde seitens der Arbeitsgruppe TELOTA der BBAW die Editionsumgebung ediarum[57] für das DTA-Basisformat angepasst. Mit ediarum können TEI-XML-basierte Editionen im Autormodus des oXygen-XML-Editors erarbeitet werden. Die Kodierung und Verwaltung der Editionsdaten erfolgt in einer eXist-Datenbank (Dumont & Fechner 2014/15). Im Rahmen des Vorhabens *Alexander von Humboldt auf Reisen* wird mit dieser Infrastruktur eine Edition nach DTA-Basisformat erstellt (Dumont et al. 2016).[58]

4.2 Nachnutzung externer wissenschaftlicher Daten im DTA

Durch die Etablierung digitaler Arbeitsmethoden steigt in den Geisteswissenschaften der Bedarf an digitalisierten Quellentexten. Dabei ist einerseits eine kritische Menge für wissenschaftlich fundierte Aussagen unerlässlich, während andererseits der Bedarf in der Datenmenge nicht zulasten der Qualität gehen darf. Was also benötigt wird, sind umfangreiche, hochwertige und standardisiert aufbereitete Textsammlungen und Korpora. Solche Daten können nur mit hohem Aufwand erzielt werden, was wiederum der Nachfrage entgegenzustehen scheint.

Auf der anderen Seite entstehen in unterschiedlichen Kontexten immer mehr solcher hochwertigen digitalen Daten, etwa im Rahmen von Editionen, in Form projekteigener oder für individuelle Forschungsarbeiten erarbeiteter Spezialkorpora oder als Textkollektionen der interessierten Community (z. B. Wikisource[59]). Gerade seitens der quantitativen Linguistik entstehen schon seit langem solcherlei digitale Forschungsdaten, die jedoch teilweise in veralteten und/oder stark individualisierten Formaten vorliegen. Die Herausforderung besteht nun darin, diese Forschungsdaten aus den verschiedenen Quellen zu ermitteln, zu standardisieren und nach einheitlichen Richtlinien aufzubereiten, sodass sie als einheitliches Gesamtkorpus auswertbar werden. Im Rahmen des Moduls DTAE verfolgt das DTA diese Aufgabe (Thomas & Wiegand 2015).[60] Dabei geht es um zweierlei: Einerseits werden Verfahren entwickelt, um Daten

56 Zugang zur KOLIMO-Plattform unter: https://kolimo.uni-goettingen.de/index.html (letzter Zugriff: 6. 11. 2017).
57 Vgl. http://www.bbaw.de/telota/software/ediarum (letzter Zugriff: 6. 11. 2017).
58 Zum Vorhaben vgl. http://avhr.bbaw.de/ (letzter Zugriff: 6. 11. 2017).
59 Vgl. https://de.wikisource.org (letzter Zugriff: 6. 11. 2017).
60 Vgl. auch die DTAE-Projektseite: http://www.deutschestextarchiv.de/dtae/ (letzter Zugriff: 6. 11. 2017).

aus größeren Textsammlungen und verschiedenen Formaten (semi-)automatisch in das DTABf zu konvertieren. Andererseits werden Wissenschaftler und Wissenschaftlerinnen angehalten und geschult, ihre individuell erstellten Daten gemäß den etablierten DTA-Vorgaben aufzubereiten und idealerweise über die DTA-Plattform auch weiteren Forscherinnen und Forschern zur Verfügung zu stellen. Über DTAE wurden bereits in über zwanzig Projekten Daten aus sehr verschiedenen Quellen und unterschiedlichen Formaten kuratiert, nicht allein größere Quellenkorpora wie die Editionstexte des Vorhabens *Johann Friedrich Blumenbach – online*[61] oder ein Korpus aus 151 Texten der Wikisource,[62] sondern auch kleinere, individuelle Datensammlungen (oft in älteren Formaten), wie etwa *Texte der deutschen Frauenbewegung*, digitalisiert von Anna Pfundt,[63] ein Korpus von Fach- und Gebrauchstexten, digitalisiert an der JLU Gießen (Thomas Gloning),[64] sowie ein historisches Zeitungskorpus, digitalisiert von Michel Lefèvre (Lefèvre 2013; in Bearbeitung).

Ein Angebot in diesem Zusammenhang ist z. B. die Möglichkeit des verteilten Arbeitens und des Crowdsourcing in DTAQ. Hier ist es möglich, als Arbeitsgruppe die eigenen Projektdaten in einem versionierten System zu verbessern und seitenbasiert tiefer zu annotieren oder auch andere Interessierte zur Mithilfe einzuladen. Für die Vorbereitung von Texten für die Integration in DTAQ wird ein Framework für den Autor-Modus des oXygen-XML-Editors bereitgestellt, das die Bearbeitung von DTABf-konformen Annotationen in einer WYSIWYG-Ansicht ermöglicht. Ein Webformular erleichtert die Erfassung von Metadaten nach DTABf. Regelmäßige Schulungen und Workshops geben Anleitung für die Anwendung der genannten Tools und die Arbeit mit dem DTA.

Die nach DTABf aufbereiteten Daten finden aus dem DTA mit automatisierten Verfahren ihren direkten Weg in die CLARIN-Infrastruktur: über CMDI-Metadatensätze im VLO, über die Anbindung aller DTA-Daten an die Föderierte Volltextsuche, über den TCF-Download sowie über die Vergabe persistenter Identifizierer und die Aufnahme in das CLARIN-Repositorium der BBAW.[65] Auf

61 Vgl. die Projektseite: http://www.blumenbach-online.de/ (letzter Zugriff: 6. 11. 2017) sowie die Beschreibung unter http://www.deutschestextarchiv.de/doku/textquellen#blumenbach (letzter Zugriff: 6. 11. 2017).

62 Vgl. die Projektseite: https://de.wikisource.org (letzter Zugriff: 6. 11. 2017) sowie die Beschreibung unter: http://www.deutschestextarchiv.de/doku/textquellen#wikisource (letzter Zugriff: 6. 11. 2017).

63 Vgl. http://www.deutschestextarchiv.de/search/metadata?corpus=tdef (letzter Zugriff: 6. 11. 2017).

64 Vgl. http://www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=tg (letzter Zugriff: 6. 11. 2017).

65 https://clarin.bbaw.de/ bzw. https://clarin.bbaw.de/de/repo/ (letzter Zugriff: 6. 11. 2017).

diese Weise wird die Nachnutzung, die Nachhaltigkeit und längerfristige Archivierung der digitalen Quellen ohne größeren Aufwand für die Datengeber gewährleistet.

Die Integration von Texten und Textsammlungen aus unterschiedlichen Quellen in das DTA trägt somit zu deren besserer Verfügbarkeit und Dissemination bei. Zudem werden die betreffenden Werke über die Plattform miteinander vernetzt, sodass sie im Zusammenhang und in Bezug zueinander recherchierbar und auswertbar sind. Dadurch können Beziehungen zwischen Werken verifiziert werden oder auch erstmals zutage treten.

Zudem konnten aus Texten unterschiedlicher Quellen Spezialkorpora gebildet werden, die das DTA-Kernkorpus sinnvoll ergänzen. So wurden im DTA Zeitungen aus bislang fünf verschiedenen Sammlungen zusammengeführt: die *Neue Rheinische Zeitung*,[66] das *Mannheimer Korpus historischer Zeitungen*,[67] der *Hamburgische Correspondent*,[68] Zeitungen aus dem Korpus der *diGiTexte*[69] sowie die Zeitschriften *Die Grenzboten* (1841–1922),[70] und J. G. Dinglers *Polytechnisches Journal* (1820–1931).[71] Des Weiteren konnte ein Korpus von Texten Alexander von Humboldts aus verschiedenen Quellen zusammengebracht werden: Humboldts

66 Digitalisiert am Rande des Vorhabens *Marx-Engels-Gesamtausgabe* (http://mega.bbaw.de/projektbeschreibung [letzter Zugriff: 6. 11. 2017]); zugänglich unter: http://www.deutschestextarchiv.de/doku/nrhz (letzter Zugriff: 6. 11. 2017).

67 Digitalisiert am Institut für Deutsche Sprache Mannheim (http://www1.ids-mannheim.de/ [letzter Zugriff: 6. 11. 2017]); zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=mkhz (letzter Zugriff: 6. 11. 2017).

68 Digitalisiert im Rahmen des Projekts *Volltextdigitalisierung der Staats- und Gelehrte[n] Zeitung des Hamburgischen Unpartheyischen Correspondenten und ihrer Vorläufer (1712–1848)* unter der Leitung von Prof. Dr. Britt-Marie Schuster an der Universität Paderborn in Zusammenarbeit mit dem DTA (https://kw.uni-paderborn.de/institut-fuer-germanistik-und-vergleichende-literaturwissenschaft/germanistische-und-allgemeine-sprachwissenschaft/schuster/forschung/projekte/der-hamburgische-unpartheyische-correspondent-volltextdigitalisierung/ [letzter Zugriff: 6. 11. 2017]). Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=correspondent (letzter Zugriff: 6. 11. 2017).

69 Digitalisiert in verschiedenen Projektkontexten an der Justus-Liebig-Universität (JLU) Gießen, Professur für Germanistische Sprachwissenschaft (Schwerpunkt Sprachverwendung), Prof. Dr. Thomas Gloning. Zugänglich unter: http://www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=tg (letzter Zugriff: 6. 11. 2017).

70 Digitalisiert im Rahmen eines DFG-Projekts der Staats- und Universitätsbibliothek (SuUB) Bremen zur Nachbearbeitung des OCR-Volltextes der Zeitschrift *Die Grenzboten*, http://brema.suub.uni-bremen.de/grenzboten (letzter Zugriff: 6. 11. 2017). Recherchierbar unter: http://kaskade.dwds.de/dstar/grenzboten/ (letzter Zugriff: 6. 11. 2017).

71 Digitalisiert im Rahmen des DFG-Projekts *Dingler Online* an der Humboldt-Universität zu Berlin (zugänglich unter: http://www.polytechnischesjournal.de/ [letzter Zugriff: 6. 11. 2017]). Recherchierbar unter: http://kaskade.dwds.de/dstar/dingler/ (letzter Zugriff: 6. 11. 2017).

unselbständige Schriften,[72] Nachschriften zu seinen Kosmos-Vorlesungen,[73] ausgewählte gedruckte Werke[74] und perspektivisch Briefe und Reisetagebücher Humboldts aus dem Projekt *Alexander von Humboldt auf Reisen*.[75] Darüber hinaus konnte eine umfangreiche Sammlung von Funeralschriften in das DTA integriert werden. Diese setzt sich aus Funeralschriften der ehemaligen Stadtbibliothek Breslau,[76] aus Epicedien Simon Dachs[77] sowie aus individuell digitalisierten Texten[78] zusammen.

Dass die geschilderten Bemühungen um die Korpora des DTA auch für die wissenschaftliche Community relevant sind, zeigt sich an der zunehmenden wissenschaftlichen Wahrnehmung und Nutzung der Korpora. So werden Ressourcen des DTA im Rahmen von Einzelstudien herangezogen oder empfohlen (z. B. Gloning 2016; Seim 2016; Schuster 2017). Darüber hinaus bauen mehrere eigene Forschungsprojekte wesentlich auf den Daten oder der Infrastruktur des DTA auf.[79]

---

72 Digitalisiert im Rahmen des Projekts *Digitalisierung ausgewählter unselbstständiger Schriften Alexander von Humboldts*, das vom DTA in Kooperation mit dem Vorhaben der BBAW *Alexander von Humboldt auf Reisen* (http://www.bbaw.de/forschung/avh-r [letzter Zugriff: 6. 11. 2017]) und der Professur für Romanische Literaturwissenschaft der Universität Potsdam (Prof. Dr. Otmar Ette) durchgeführt wurde. Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=avh (letzter Zugriff: 6. 11. 2017).

73 Digitalisiert im Rahmen einer Kooperation des DTA mit dem Projekt *Hidden Kosmos* (siehe auch oben, Abschnitt 2.4). Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=avhkv (letzter Zugriff: 6. 11. 2017).

74 Digitalisiert für das DTA-Kernkorpus; zugänglich unter: http://www.deutschestextarchiv.de/api/pnd/118554700 (letzter Zugriff: 6. 11. 2017).

75 In Vorbereitung, vgl. http://edition-humboldt.de/ (letzter Zugriff: 24. 11. 2017).

76 Digitalisiert im Rahmen des DFG-Projekts *AEDit Frühe Neuzeit* der Herzog August Bibliothek Wolfenbüttel (HAB), des DTA an der BBAW und der Forschungsstelle für Personalschriften an der Philipps-Universität Marburg (http://diglib.hab.de/?link=029 [letzter Zugriff: 6. 11. 2017]). Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=aedit (letzter Zugriff: 6. 11. 2017).

77 Digitalisiert im Rahmen des DFG-Pilotprojektes zum *OCR-Einsatz bei der Digitalisierung der Funeralschriften der Staatsbibliothek zu Berlin*; Federbusch & Polzin 2013; zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=sbb_funeralschriften (letzter Zugriff: 6. 11. 2017).

78 Digitalisiert im Rahmen von Projektarbeiten im Rahmen von Seminaren zu Methoden der digitalen Textedition an der Freien Universität zu Berlin (Dozenten: Matthias Boenig, Susanne Haaf).

79 Aktuell sind hierbei z. B. die Projekte *Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch* (http://gepris.dfg.de/gepris/projekt/279165027 [letzter Zugriff: 6. 11. 2017]); *Redewiedergabe – Eine literatur- und sprachwissenschafliche Korpusanalyse* (http://gepris.dfg.de/gepris/projekt/322751860) und *Digitale Sammlung Deutscher Kolonialismus* (bewilligt im März 2017) zu nennen.

5 Zusammenfassung und Ausblick

Das Deutsche Textarchiv wurde von 2007 bis 2016 von der Deutschen Forschungsgemeinschaft gefördert. In dieser Zeit wurde ein nach Textsorten und über die Zeit ausgewogenes Kernkorpus von Texten aus der Zeit von etwa 1600 bis 1900 im Umfang von ca. 120 Millionen Textwörtern aufgebaut. Darüber hinaus entwickelte sich das DTA zu einer von vielen Textproduzenten genutzten Korpusplattform. Zwischen 2011 und 2016 konnten Kooperationen mit über zwanzig institutionell geförderten Korpusprojekten vereinbart werden. Hierdurch entstanden zahlreiche weitere Texte für das DTA. Insgesamt stehen somit im DTA insgesamt mehr als 1 Million Seiten hochqualitativer deutschsprachiger Texte zur Verfügung. Das DTA hat darüber hinaus eine Vielzahl von Nachnutzungen erfahren. Allein zwischen Juli 2015 und April 2017 wurde das DTA als „Gesamtpaket“ mehr als 5.000-mal heruntergeladen und zu Forschungszwecken oder für die universitäre Lehre eingesetzt. Etwa 50 Nachnutzungen des Korpus führten zu Veröffentlichungen im Kontext der Digital Humanities.[80]

Im Dezember 2016 lief die Förderung des Deutschen Textarchivs durch die DFG aus. Durch die Aufnahme in den CLARIN-Verbund seit 2014 können technische Softwareentwicklungen und auch der Aufbau eigener Korpora zwar nicht mehr in dem Maße wahrgenommenen werden wie in der durch die DFG geförderten Aufbauphase. Seinen zentralen Aufgaben kann das DTA jedoch durch die Integration in das CLARIN-Zentrum weiter nachkommen. Die BBAW ist seit 2016 im deutschen Teil von CLARIN, dem Projekt CLARIN-D, Koordinator des Kompetenzbereichs *Historische Daten* und kann in diesem Rahmen zentralen Aufgaben der Nutzung des Deutschen Textarchivs nachkommen. Diese bestehen im weiteren zuverlässigen Betrieb der Plattformen DTA und DTAQ, der Weiterentwicklung bzw. Wartung des interoperablen TEI-kompatiblen Schemas DTABf sowie in der nachhaltigen Aufbewahrung der Korpusressourcen und der Services des DTA. Im Rahmen von CLARIN-D engagiert sich das DTA nach wie vor, um durch Kooperationen mit Korpusaufbauprojekten die Textbasis des DTA zu vergrößern. Zusammengefasst kann somit festgestellt werden, dass das DTA das selbstgesteckte Ziel, als aktives Archiv zu fungieren, erfüllt und damit als wichtiger Bestandteil im „Ökosystem“ der Digital Humanities verankert ist.

80 Vgl. http://www.deutschestextarchiv.de/clarin-kooperationen (letzter Zugriff: 6. 11. 2017).

Literatur

Bird, Christian, Tim Menzies & Thomas Zimmermann (2015): *The art and science of analyzing software data*. San Francisco, CA: Morgan Kaufmann.

DFG (2015a): *Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*. Hrsg. vom Fachkollegium Sprachwissenschaften der Deutschen Forschungsgemeinschaft (DFG). http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (letzter Zugriff: 6. 11. 2017).

DFG (2015b): *Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft*. Hrsg. vom Fachkollegium Literaturwissenschaft der DFG. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf (letzter Zugriff: 6. 11. 2017).

Dumont, Stefan & Martin Fechner (2014/15): Bridging the gap: Greater usability for TEI encoding. *Journal of the Text Encoding Initiative [Online]* 8. http://jtei.revues.org/1242 (letzter Zugriff: 6. 11. 2017).

Dumont, Stefan, Susanne Haaf, Tobias Kraft, Alexander Czmiel, Christian Thomas & Matthias Boenig (2016): Applying standard formats and tools, 69 f. *TEI Conference and Members Meeting 2016: Book of abstracts*. http://tei2016.acdh.oeaw.ac.at/sites/default/files/TEIconf2016_BookOfAbstracts.pdf#page=71 (letzter Zugriff: 6. 11. 2017).

Eckart, Kerstin (2012): Resource annotations. Aspects of annotations. *CLARIN-D User Guide*. Chapter I,3[,1]. http://media.dwds.de/clarin/userguide/text/annotation_aspects.xhtml.

Federbusch, Maria & Christian Polzin (2013): *Volltext via OCR – Möglichkeiten und Grenzen. Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz. Mit einem Erfahrungsbericht von Thomas Stäcker aus dem Projekt „Helmstedter Drucke Online“ der Herzog August Bibliothek Wolfenbüttel*. Berlin: o. V. (= Beiträge aus der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz 43.).

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas & Frank Wiegand (2012): TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. *Jahrbuch für Computerphilologie – online*. http://computerphilologie.digital-humanities.de/jg09/geykenetal.html (letzter Zugriff: 6. 11. 2017).

Geyken Alexander, Susanne Haaf & Frank Wiegand (2012): The DTA ‘base format’. A TEI-subset for the compilation of interoperable corpora. In Jeremy Jancsary (Hrsg.), *Proceedings of the 11th Conference on Natural Language Processing (KONVENS)*, 383–391. Vienna: Eigenverlag ÖGAI (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5). http://www.oegai.at/konvens2012/proceedings.pdf#page=383 (letzter Zugriff: 6. 11. 2017).

Geyken, Alexander, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, Frank Wiegand & Kay-Michael Würzner (2015): Zeitliche Verlaufskurven in den DTA- und DWDS-Korpora: Wörter und Wortverbindungen über 400 Jahre (1600–2000). In *DHd 2015: Von Daten zu Erkenntnissen: Book of Abstracts*, 78–88. http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege#page=78 (letzter Zugriff: 6. 11. 2017).

Gloning, Thomas (2016): Kommunikationsgeschichte, Themengeschichte, Ideengeschichte. Beispiele für Zusammenhänge und Lehrszenarien. In Volker Harm, Holger Runow & Leevke Schiwek (Hrsg.), *Sprachgeschichte des Deutschen. Positionierungen in Forschung, Studium, Unterricht*, 181–201. Stuttgart: Hirzel.

Haaf, Susanne (2017): Das DTA-Basisformat in neuem Gewand. In *Im Zentrum Sprache. Untersuchungen zur deutschen Sprache*, 3. März 2017. https://sprache.hypotheses.org/147 (letzter Zugriff: 6. 11. 2017).

Haaf, Susanne & Matthias Schulz (2014): Historical newspapers & journals for the DTA. In: *Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage (LRT4HDA). Proceedings of the workshop, held at the 9th LREC, May 26–31, Reykjavik (Iceland)*, 50–54. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRT4HDA%20Proceedings.pdf#page=57 (letzter Zugriff: 6. 11. 2017).

Haaf, Susanne, Alexander Geyken & Frank Wiegand (2014/2015): The DTA 'base format': A TEI subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative 8*. http://jtei.revues.org/1114 (letzter Zugriff: 6. 11. 2017).

Haaf, Susanne & Christian Thomas (2016): Die Historischen Korpora des Deutschen Textarchivs als Grundlage für sprachgeschichtliche Forschungen. In Volker Harm, Holger Runow & Leevke Schiwek (Hrsg.), *Sprachgeschichte des Deutschen. Positionierungen in Forschung, Studium, Unterricht*, 217–234. Stuttgart: Hirzel.

Haaf, Susanne & Christian Thomas (2016 [2017]): Enabling the encoding of manuscripts within the DTABf: Extension and modularization of the format. *Journal of the Text Encoding Initiative (jTEI) 10: Conference Issue*. DOI: 10.4000/jtei.1650. http://jtei.revues.org/1650 (letzter Zugriff: 6. 11. 2017).

Hahn, Udo, Franz Matthies, Erik Faessler & Johannes Hellrich (2016): Uima-based JCoRe 2.0 goes GitHub and Maven Central: State-of-the-art software resource engineering and distribution of NLP pipelines. *Proceedings of the 10th LREC 2016*, 2502–2509. http://www.lrecconf.org/proceedings/lrec2016/pdf/774_Paper.pdf (letzter Zugriff: 6. 11. 2017).

Hamp, Birgit & Helmut Feldweg (1997): GermaNet – a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9–15. Madrid: o. V. http://www.aclweb.org/anthology/W97-0802 (letzter Zugriff: 6. 11. 2017).

Heid, Ulrich, Helmut Schmid, Kerstin Eckart & Erhard Hinrichs (2010): A corpus representation format for linguistic web services: The D-SPIN text corpus format and its relationship with ISO standards. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*. Malta: o. V.

Hellrich, Johannes, Franz Matthies & Udo Hahn (2017): UIMA als Plattform für die nachhaltige Software-Entwicklung in den Digital Humanities. In *Dhd 2017: Digitale Nachhaltigkeit: Book of Abstracts*, 279–281. http://www.dhd2017.ch/wp-content/uploads/2017/03/Abstractband_def3_M%C3%A4rz.pdf (letzter Zugriff: 24. 11. 2007).

Henrich, Verena & Erhard Hinrichs (2010). GernEdiT – The GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC '10)*, 2228–2235. Malta: o. V. http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf (letzter Zugriff: 6. 11. 2017).

Hinrichs, Erhard, Marie Hinrichs & Thomas Zastrow (2010): WebLicht. Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, 25–29.

Jurish, Bryan (2012): Finite-state canonicalization techniques for historical German. Dissertation. Potsdam: Universität Potsdam. urn:nbn:de:kobv:517-opus-55789. http://opus.kobv.de/ubp/volltexte/2012/5578/ (letzter Zugriff: 6. 11. 2017).

Jurish, Bryan & Kay-Michael Würzner (2013): Word and sentence tokenization with Hidden Markov Models. *Journal for Language Technology and Computational Linguistics* 28(2), 61–83.

Jurish, Bryan, Christian Thomas & Frank Wiegand (2014): Querying the Deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, & C. Gurrin (Hrsg.), *Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities*, 25–30. urn:nbn:de:0074-1131-6. http://ceur-ws.org/Vol-1131/ (letzter Zugriff: 24. 11. 2017).

Jurish, Bryan, Alexander Geyken & Thomas Werneke (2016): DiaCollo: diachronen Kollokationen auf der Spur. In *DHd 2016: Modellierung – Vernetzung – Visualisierung: Book of Abstracts*, 172–175. Duisburg: nisaba verlag.

Jurish, Bryan (2015): DiaCollo: On the trail of diachronic collocations. In K. De Smedt (Hrsg.), *Proceedings of the CLARIN Annual Conference 2015*, 28–31. http://www.deutschestextarchiv.de/files/jurish2015diacollo-clarin.pdf (letzter Zugriff: 6. 11. 2017).

Kuczera, Andreas (2017): Das Deutsche Textarchiv in der Graphenwelt. In *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*. https://mittelalter.hypotheses.org/10025 (letzter Zugriff: 6. 11. 2017).

Lefèvre, Michel (2013): *Textgestaltung, Äußerungsstruktur und Syntax in deutschen Zeitungen des 17. Jahrhunderts. Zwischen barocker Polyphonie und solistischem Journalismus*. Berlin: Weidler.

Lemnitzer, Lothar, Bryan Jurish und Daniel Burkhardt (2016): *DiaCollo Tutorial*. http://kaskade.dwds.de/diacollo-tutorial (letzter Zugriff: 6. 11. 2017).

Seim, Stefanie (2016): *Nominalphrasen in literarischen Texten: Strukturtypen und Funktionen beim Figurenentwurf in Werken des 20. und 21. Jahrhunderts*. Gießen: Gießener Elektronische Bibliothek (= Linguistische Untersuchungen 10).

Sinclair, Stéfan & Geoffrey Rockwell (2017): *Voyant Tools*. http://voyant-tools.org/ (letzter Zugriff: 6. 11. 2017).

Thomas, Christian & Frank Wiegand (2015): Making great work even better. Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th–19th centuries) in CLARIN-D. In Jost Gippert & Ralf Gehrke (Hrsg.), *Historical corpora. Challenges and perspectives*, 181–196. Tübingen: Narr.

Tiepmar, Jochen, Thomas Eckart, Dirk Goldhahn & Christoph Kuras (2016): Canonical text services in CLARIN – Reaching out to the Digital Classics and beyond. In *CLARIN Annual Conference*. http://cts.informatik.uni-leipzig.de/documents/CLARIN_CTS.pdf (letzter Zugriff: 6. 11. 2017).

Tiepmar, Jochen, Thomas Eckart, Dirk Goldhahn & Christoph Kuras (2017): Integrating canonical text services into CLARIN's search infrastructure. In: *Linguistics and Literature Studies* 5, 99–104. doi:10.13189/lls.2017.050205, http://www.hrpub.org/journals/article_info.php?aid=5738 (letzter Zugriff: 6. 11. 2017).

Wittenburg, Peter & Dieter van Uytvanck (2012): Metadata. The Component Metadata Initiative (CMDI). In *CLARIN-D User Guide*, Chapter I,2[,6] sowie dies. Metadata. Aggregation. Ebd., Chapter I,2[,7]. http://media.dwds.de/clarin/userguide/text/metadata_aggregation.xhtml (letzter Zugriff: 6. 11. 2017).