
**Susanne Haaf, Christian Thomas**

# Die historischen Korpora des Deutschen Textarchivs als Grundlage für sprachgeschichtliche Forschungen

PREPRINT-Version 2016

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-25112

Die vorliegende Datei wird Ihnen von der Berlin-Brandenburgischen Akademie der Wissenschaften unter einer Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Germany (cc by-nc-sa 3.0) Licence zur Verfügung gestellt.

– PREPRINT-Version, erscheint (2016) in: Holger Runow/Volker Harm/Levke Schiwek (Hgg.): *Sprachgeschichte des Deutschen: Positionierungen in Forschung, Studium, Schule*. Stuttgart: Hirzel. –

# DIE HISTORISCHEN KORPORA DES DEUTSCHEN TEXTARCHIVS ALS GRUNDLAGE FÜR SPRACHGESCHICHTLICHE FORSCHUNGEN

*Susanne Haaf, Christian Thomas*

## I. EINFÜHRUNG

Das an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) beheimatete DFG-geförderte Projekt Deutsches Textarchiv (DTA)[1] erarbeitet ein disziplinenübergreifendes Volltextkorpus deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis ca. 1900. Die Korpuslandschaft für das historische Neuhochdeutsche war zuvor eher dünn besiedelt und konnte korpuslinguistischen Forschungsfragen nur bedingt gerecht werden. Zwar existieren große Textsammlungen historischer Werke im Internet, wie z. B. zeno.org[2], Wikisource[3] oder die Volltexte von Google Books[4], doch folgen diese Sammlungen in ihrer Zusammenstellung, der Auswahl der Vorlagen und der Qualität der präsentierten Texte nicht konsequent wissenschaftlichen Kriterien und sind daher aus unterschiedlichen Gründen für die Wissenschaft nur bedingt nachnutzbar. So wurden etwa die Texte von zeno.org in der Regel gegenüber den ursprünglichen Ausgaben in ihrer Schreibung modernisiert.[5] Die von Google Books bereitgestellten OCR-Volltexte sind stark fehlerbehaftet, häufig mit ebenso fehlerhaften Metadaten versehen, die Textauswahl erfolgte ungerichtet, Dopplungen sind dabei ebenfalls möglich etc. Das Projekt Wikisource kommt im Vergleich dazu einem wissenschaftlichen Interesse schon näher. Zwar erfolgt auch hier die Textauswahl ungerichtet, jedoch ist durch die Methode der manuellen Transkription mit mehrfachen Nachkorrekturen die Erfassungsqualität sehr hoch. Allerdings folgt die Textauszeichnung nicht den verbreiteten Standards, die Werke sind nicht tiefer linguistisch erschlossen und komplexe Suchanfragen über das Korpus nicht möglich. Neben diesen großen

1 URL: *www.deutschestextarchiv.de* [wie alle URL in diesem Dokument zuletzt abgerufen am 14.7.2014].
2 Zeno.org-Volltextbibliothek. URL: *http://www.zeno.org*.
3 Wikisource, die freie Quellensammlung. URL: *https://de.wikisource.org*.
4 Google Books. URL: *http://books.google.de*.
5 Vgl. dazu auch Thomas/Wiegand 2015.

Volltextsammlungen existieren eher kleinere, meist aus thematisch fokussierten Editionsprojekten sich ergebende Textsammlungen.

Vor diesem Hintergrund verfolgt das Projekt Deutsches Textarchiv das Ziel, die Grundlage für ein Referenzkorpus des historischen Neuhochdeutschen zu erarbeiten und damit eine Korpusbasis für linguistische Untersuchungen zu dieser Sprachstufe des Deutschen zu schaffen. Ein grundlegendes Konzept dabei ist, historische Werke jeweils vollständig (im Gegensatz zur auszugsweisen Erfassung einzelner Abschnitte oder Seiten) zu erfassen. Damit können die im DTA zur Verfügung gestellten Volltexte auch von verschiedenen anderen Disziplinen jenseits der Linguistik nachgenutzt werden, beispielsweise als Basistexte für die Editionswissenschaft, als leicht zugängliche Quellen für Historiker und Wissenschaftshistoriker oder als Materialien für eine neue Form des Schulunterrichts, welcher digitale Ressourcen einbezieht.

## II. DAS PROJEKT DEUTSCHES TEXTARCHIV (DTA)

Ziel des DTA ist es, einen vielseitigen, umfangreichen Textbestand zu verschiedenen Varietäten des Deutschen bereitzustellen, anhand dessen die Entwicklung der neuhochdeutschen Sprache vom 17. bis zum 19. Jahrhundert nachvollzogen werden kann. Auf der Grundlage digitaler Faksimiles werden die jeweiligen historischen Druckwerke vollständig erfasst, in einem flexiblen, interoperablen XML-Format annotiert und linguistisch aufbereitet.

Das DTA-Kernkorpus enthält zum einen wohlbekannte und weitreichend rezipierte Werke – beispielsweise aus dem Bereich der ‚schönen' Literatur Friedrich Schillers Drama ‚Kabale und Liebe' (Schiller 1784) sowie Goethes zweibändigen ‚Werther'-Roman (Goethe 1774), oder, als ein Beispiel aus dem Bereich der Gebrauchsliteratur, das heute synonym mit seinem Verfasser einfach ‚Knigge' genannte Werk ‚Ueber den Umgang mit Menschen' (2 Bde., Knigge 1788). Hinzu kommen zahlreiche Werke, die als prägend für die Geschichte eines bestimmten Fachbereichs bzw. der Wissenschaft(ssprache) im Allgemeinen gelten können, wie etwa Justus von Liebigs Arbeiten über organische Chemie (2 Bde., Liebig 1840 und 1842) oder Wilhelm Conrad Röntgens in drei Teilen erschienene Mitteilungen über die von ihm so genannten „X-Strahlen" (Röntgen 1896a, 1896b und 1897). Zum anderen finden sich im DTA etliche Werke, die für ihre Zeit bedeutend und mutmaßlich prägend für den Sprachgebrauch waren, heutzutage aber eher in Vergessenheit geraten sind, etwa Christian Fürchtegott Gellerts Roman ‚Das Leben der Schwedischen Gräfinn von G.***' (2 Bde., Gellert 1747 und 1748) oder die ‚Auserlesene[n] Gedichte' von Anna Luise Karsch (1764). Dieser Kernbestand des DTA wird nicht allein durch eigene Volltextdigitalisierungen weiterer Werke, sondern auch im Rahmen verschiedener Kooperationsprojekte sowie durch Kuration und Integration geeigneter externer Textressourcen systematisch um historische Primärtexte ergänzt. Das resultierende Gesamtkorpus aus DTA-Kernkorpus und seinen Erweiterungen umfasst derzeit (Stand: Juli 2014)

mehr als 750 000 digitalisierte Seiten, die der textbasierten Forschung zur vielfältigen Nutzung zur Verfügung stehen.

Im vorliegenden Beitrag werden zunächst die Prinzipien und Methoden des Korpusaufbaus im DTA vorgestellt. In Bezug auf das DTA-Kernkorpus stehen dabei zunächst die Prinzipien für die Textauswahl, die Auswahl der Digitalisierungsvorlagen und die Bilddigitalisierung sowie der Prozess der Volltexterstellung (Transkription) im Vordergrund. Aber auch für die Einbindung externer Textressourcen in das DTA existiert ein mittlerweile etablierter Workflow. Im Anschluss an die Erläuterung dieser Präliminarien werden die Verfahren und Richtlinien des DTA, welche für alle DTA-Texte (Kernkorpus und Erweiterungskorpus) gleichermaßen gelten, von der Textstrukturierung über die automatisierte linguistische Analyse bis hin zu einer umfassenden, webbasierten Qualitätssicherung vorgestellt. Abschließend werden die Möglichkeiten der Arbeit mit der DTA-Infrastruktur sowie den DTA-Korpora beispielhaft gezeigt und damit Anregungen gegeben für den Einsatz der DTA-Korpora nicht allein für sprachhistorische Forschungen, sondern auch zur Vermittlung sprach- und kulturgeschichtlicher Inhalte im universitären und schulischen Unterricht.

## III. AUFBAU UND ERSCHLIESSUNG DER KORPORA DES DEUTSCHEN TEXTARCHIVS

### I.1 DAS DTA-KERNKORPUS

*a) Textauswahl*. Der Textauswahl für das DTA liegt eine umfassende, durch Wissenschaftlerinnen und Wissenschaftler verschiedener Disziplinen getroffene und ergänzte Textauswahl zugrunde. Die Textauswahl umfasst Werke aller belletristischen Genres ebenso wie Sachbücher und wissenschaftliche Texte zu verschiedenen Themen, Wissens- und Lebensbereichen. Mit dieser Vielfalt an ausgewählten Texten soll ein Querschnitt durch das sprachliche Inventar des Deutschen über den gesamten Zeitraum des historischen Neuhochdeutschen (von ca. 1600 bis ca. 1900) erreicht werden.[6] Um möglichst authentische Ergebnisse zum Zustand und zur Entwicklung des Sprachgebrauchs zu ermöglichen, werden der Digitalisierung nach Möglichkeit die Erstausgaben der jeweiligen Werke zugrunde gelegt und wird deren Text vorlagengetreu ohne ‚Normalisierungen' erfasst. Das DTA-Kernkorpus umfasst derzeit (Juli 2014) rund 1300 Texte, weitere 200 Werke befinden sich in Bearbeitung. Auch wenn das DTA-Korpus damit bereits jetzt sehr umfangreich ist und durch Eigendigitalisierungen weiter wächst, sind darüber hinaus sukzessive Ergänzungen um weitere Werke wünschenswert, um die Datenbasis für Fragestellungen an das Korpus noch zu vergrößern. Solche Ergänzungstexte werden im Rahmen des Moduls DTAE (DTA-Erweiterungen)

[6] Siehe dazu ausführlicher *www.deutschestextarchiv.de/doku/textauswahl*.

aus anderen Projektkontexten übernommen.[7] Auf diese Weise konnten bislang weitere 700 Werke kuratiert, in homogener Weise entsprechend den DTA-Richtlinien aufbereitet und in das DTA-Korpus integriert werden.

*b) Bilddigitalisierung.* Sämtlichen Werken des DTA liegen Bilddigitalisate der jeweiligen Ausgaben zugrunde, welche in der Regel von den besitzenden Bibliotheken selbst angefertigt werden bzw. als Resultat größerer Bilddigitalisierungsprojekte[8] bereits vorliegen. Dabei ist eine ausgezeichnete Bildqualität mit möglichst hoher Auflösung unbedingt notwendig, da sich Abstriche in diesem Bereich rasch auf die Erfassungsgenauigkeit auswirken und nicht mehr problemlos kompensiert werden können. Die Faksimiles werden immer zusammen mit der Textausgabe publiziert, sodass sämtliche Transkriptionen und Annotationen jederzeit am Original überprüfbar sind.

*c) Die Transkription und ihre Richtlinien.* Die Transkription der historischen Drucke erfolgt in Zusammenarbeit mit einem Dienstleister im Double Keying-Verfahren durch Nicht-Muttersprachler.[9] Bei diesem Verfahren werden die Texte von zwei getrennt voneinander arbeitenden Personen manuell erfasst; beide daraus resultierenden Fassungen werden anschließend mit Hilfe einer Software verglichen. Die dabei sichtbar werdenden Differenzen beider Transkriptionen werden nochmals mit der Vorlage abgeglichen und zugunsten der korrekten Wiedergabe des Textes vereinheitlicht. Die resultierenden Transkriptionen erhalten dadurch einen außerordentlich hohen Grad an Genauigkeit auf der Zeichenebene (Haaf/Wiegand/Geyken 2013). Strukturierende (formale) Merkmale der Vorlage (z. B. Kopf- und Fußzeilen, Seiten- und Zeilenumbrüche, Abbildungen oder typographische Hervorhebungen) sowie inhaltliche Merkmale (z. B. die Kapitelstrukturierung, Strophen und Verse bei Gedichten, Sprechakte und Bühnenanweisungen im Drama) werden ebenfalls bereits im Zuge der Texterfassung gekennzeichnet. Diese Kennzeichnungen erfolgen zunächst in einem vereinfachten XML-Format. In einem zweiten Schritt werden diese vereinfachten XML-Auszeichnungen teilautomatisch in das TEI/XML-basierte DTA-Basisformat (DTABf) konvertiert.[10]

7 Siehe dazu unten Kap. III.2: Das DTA-Erweiterungskorpus.

8 Z. B. die im Zusammenhang mit den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke (VD) des 16., 17. bzw. 18. Jahrhundert digitalisierten Werke. Vgl. dazu die einzelnen Webseiten des VD 16, *www.vd16.de*, VD 17, *www.vd17.de* bzw. VD 18, *http://vd18.de/*.

9 Lediglich ein kleiner Teil von einfach strukturierten, nach ca. 1850 gedruckten Texten wurde zunächst durch automatische Zeichenerkennung (*Optical Character Recognition, OCR*) erfasst und anschließend intensiv semiautomatisch korrigiert, um auch hier eine ausreichend hohe Zeichengenauigkeit sicherzustellen. Siehe dazu auch die Grafiken zur Verteilung der Werke/Seiten aus dem DTA-Kernkorpus nach Digitalisierungsmethode unter *www.deutschestextarchiv.de/dtaq/stat/digimethod*.

10 Siehe dazu unten Kap. III.3: Textstrukturierung nach DTA-Basisformat (DTABf).

## III.2 DAS DTA-ERWEITERUNGSKORPUS (DTAE)

Der DTA-Kernbestand aus insgesamt 1500 ausgewählten Werken wird, wie bereits erwähnt, zusätzlich erweitert durch Primärtexte zur Sprach- und Kulturgeschichte des Deutschen, die in anderen Kontexten digitalisiert wurden. Oft finden sich wertvolle Transkriptionen an entlegenen Stellen im Netz verstreut oder werden auf lokalen Datenträgern in proprietären, veralteten und unflexiblen Formaten vorgehalten. Die Auffindbarkeit, Nachnutzung sowie weitere Bearbeitung dieser Textressourcen ist somit erheblich erschwert; viele dieser Ressourcen sind der Forschung bislang kaum bekannt.

Das DTA verfügt mit seinem Erweiterungsmodul DTAE[11] über ein Software-Paket und einen etablierten Workflow, um ausgewählte Texte dieser Art als Subkorpustexte in das DTA zu integrieren. Ein zentraler Arbeitsschritt ist dabei die Konvertierung der externen Texte aus den verschiedenen Formaten in das einheitliche DTA-Basisformat, welches sämtlichen DTA-Korpustexten als Auszeichnungsformat zugrunde liegt. Anschließend erfolgt die Integration in das unten beschriebene Qualitätssicherungsmodul DTAQ, wodurch die qualitative Anpassung der Texte an das Niveau der DTA-Korpora ermöglicht wird. Schließlich werden auch diese externen Texte mithilfe der im DTA entwickelten und ständig verfeinerten Tools einheitlich linguistisch erschlossen. Durch diese Arbeitsschritte werden externe Texte entsprechend den DTA-Richtlinien so aufbereitet, dass ein hinsichtlich der Transkriptionsgenauigkeit, der Textstrukturierung und der linguistischen Annotation homogenes Textkorpus entsteht, welches nachhaltig, interoperabel und somit vielfältig nachnutzbar ist.

## III.3 TEXTSTRUKTURIERUNG NACH DTA-BASISFORMAT (DTABf)

Wie bereits erwähnt, wird für die einheitliche Textstrukturierung innerhalb der DTA-Korpora das DTA-Basisformat (DTABf) zugrunde gelegt. Das DTABf stellt eine echte Teilmenge des Tagsets der *Text Encoding Initiative* (TEI) in seiner aktuellen Fassung TEI-P5 dar.[12] Die TEI-Richtlinien haben sich mittlerweile international als *de facto*-Standard für die Repräsentation elektronischer Texte etabliert.

Ziel der Reduktion des sehr umfangreichen TEI-Tagsets im DTABf ist es, orientiert an den Gegebenheiten historischer Druckvorlagen, wie sie im DTA bearbeitet werden, und den daraus abgeleiteten Erfordernissen für deren Aufbereitung, eine besser überschaubare Auswahl an Kodierungsmöglichkeiten zu bilden. Diese sollen zudem genau eine eindeutige Lösung für die Annotation des jeweiligen Phänomens bieten, um eine einheitliche Auszeichnung des Korpus zu gewährleisten. Das folgende Beispiel veranschaulicht diese Bemühungen:

[11] *www.deutschestextarchiv.de/dtae*.
[12] *http://tei-c.org; http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html*.

TEI-P5-konformes Tagging von Eigennamen:

```
<rs type="propNounPersName">Sokrates</rs>,
<name type="person">Platon</name> und
<persName>Aristoteles</persName> waren Philosophen.
```

DTABf- und TEI-P5-konformes Tagging von Eigennamen:

```
<persName>Sokrates</persName>,
<persName>Platon</persName> und
<persName>Aristoteles</persName> waren Philosophen.
```

*Beispiel: Reduktion des TEI-Tagsets im DTA-Basisformat.*

Das Beispiel zeigt, dass die TEI-Richtlinien mehrere Möglichkeiten der Annotation von Eigennamen zulassen. Dies kann zu Inkohärenzen im Tagging eines Textes und folglich zu Schwierigkeiten bei der Arbeit mit den so annotierten Texten führen. Die TEI adressiert dieses Problem und ermöglicht mit dem ODD-Format die Reduktion des TEI-Tagsets für den konkreten Anwendungsfall.[13] Im DTABf wurde diese Möglichkeit genutzt, um z. B. das TEI-P5-Tagset auf eine Lösung für die Eigennamenauszeichnung zu reduzieren und so Inkohärenzen bei der Textauszeichnung zu vermeiden.
Das DTABf sieht die formale ebenso wie die inhaltliche Strukturierung der Texte vor (s. Abb. 1). Es wurde aufgrund der in den Korpustexten des DTA beobachteten Phänomene entwickelt und ausführlich mit konkreten Beispielen aus dem Korpus dokumentiert. Ein RNG-Schema und ergänzende Schematron-Regeln ermöglichen darüber hinaus die formale Validierung der TEI-Dokumente gegen das DTABf. Beide Komponenten des DTABf (Dokumentation einerseits, formale Spezifikation mittels Schema andererseits) gewährleisten eine vollständig DTABf-konforme Textauszeichnung.[14]

[13] Vgl. TEI Consortium (2014: 22; 23.3).

[14] Die Dokumentation, das RNG-Schema und der Schematron-Regelsatz des DTABf sind zugänglich unter: *www.deutschestextarchiv.de/doku/basisformat*. Zum DTABf vgl. auch Geyken/Haaf/Wiegand 2012.

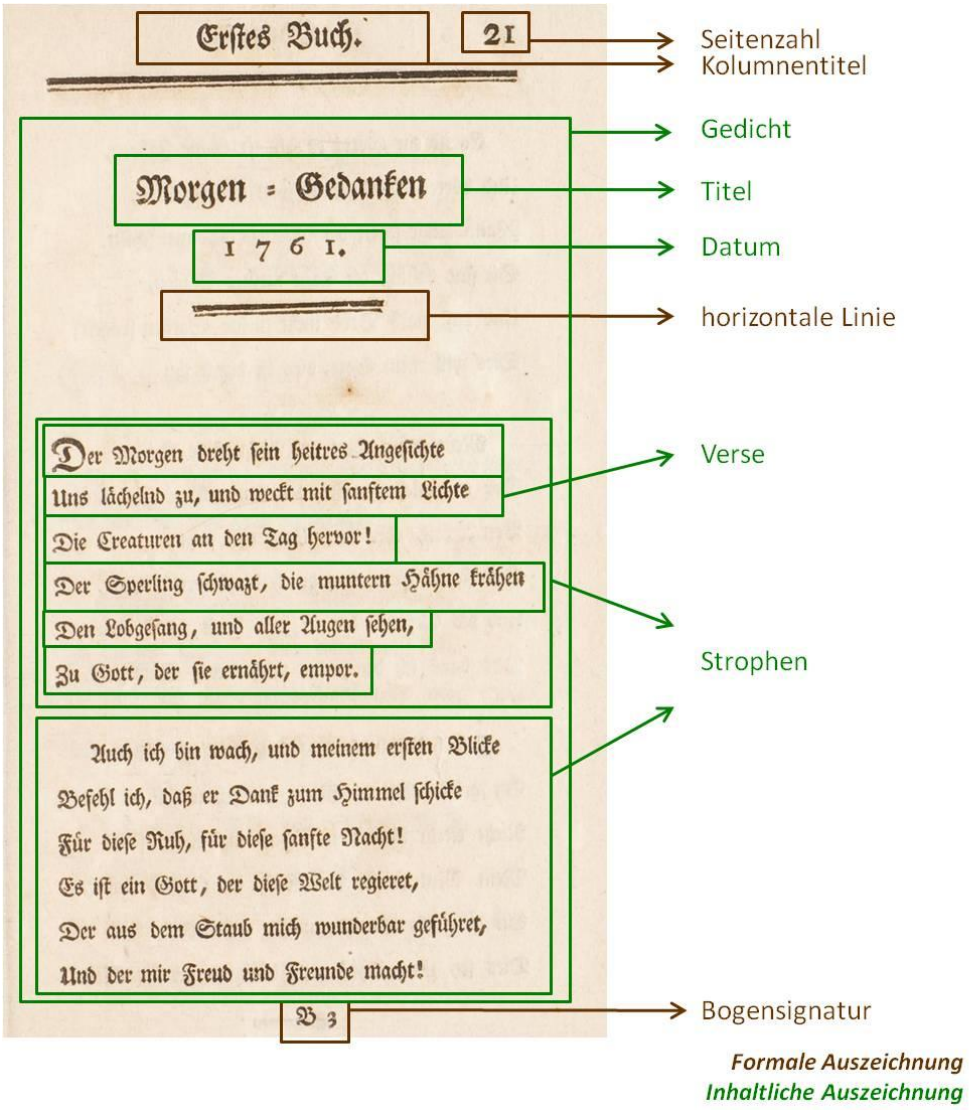



*Abb. 1: Auszeichnung formaler und inhaltlicher Strukturelemente innerhalb des DTA; hier: Karsch 1764, S. 21.*

Durch die grundsätzliche Prämisse, für gleichartige Phänomene nur genau eine Kodierungsmöglichkeit zu bieten, gewährleistet das DTABf die homogene Textauszeichnung für sämtliche DTA-Korpustexte. Es wird also nicht allein für das DTA-Kernkorpus, sondern auch für die im Rahmen von Kooperations- und Kurationsprojekten in das DTA integrierten Texte eingesetzt. Dabei wurden im DTA verschiedene Workflows entwickelt, um Texte aus unterschiedlichen Formaten (z. B. Word, Wiki-Syntax, andere TEI-Dialekte) in das DTABf zu überführen.[15]

[15] Wenngleich viele der diesbezüglichen Arbeitsschritte automatisiert werden konnten, ist in der Regel zusätzlich ein – je nach Vorlage mehr oder weniger großer – manueller Aufwand von-


*Abb. 2: Strukturierung in TEI-XML nach DTABf, hier: Wedekind 1891, Titelseite (XML-Ansicht).*

Ein besonderer Fokus liegt darüber hinaus auf der Bereitstellung ausführlicher Metadaten, in welchen Titelinformationen, Verantwortlichkeiten, Nachnutzungsrechte, Angaben zur physischen Quelle, Hinweise auf die zugrundeliegenden Transkriptions- und Annotationsrichtlinien sowie erste inhaltliche Informationen verzeichnet werden. Die ausführlichen Metadaten werden in verschiedene Metadatenformate konvertiert und können somit vielfältig nachgenutzt und in unterschiedlichen Portalen verfügbar gemacht werden.

nöten, um die vollständig DTABf-konforme Strukturierung sämtlicher DTA-Textdaten bis hin zu einem einheitlichen Strukturierungslevel und somit eine einheitlich hohe Qualität bei der Textauszeichnung für sämtliche DTA-Korpustexte zu erreichen.

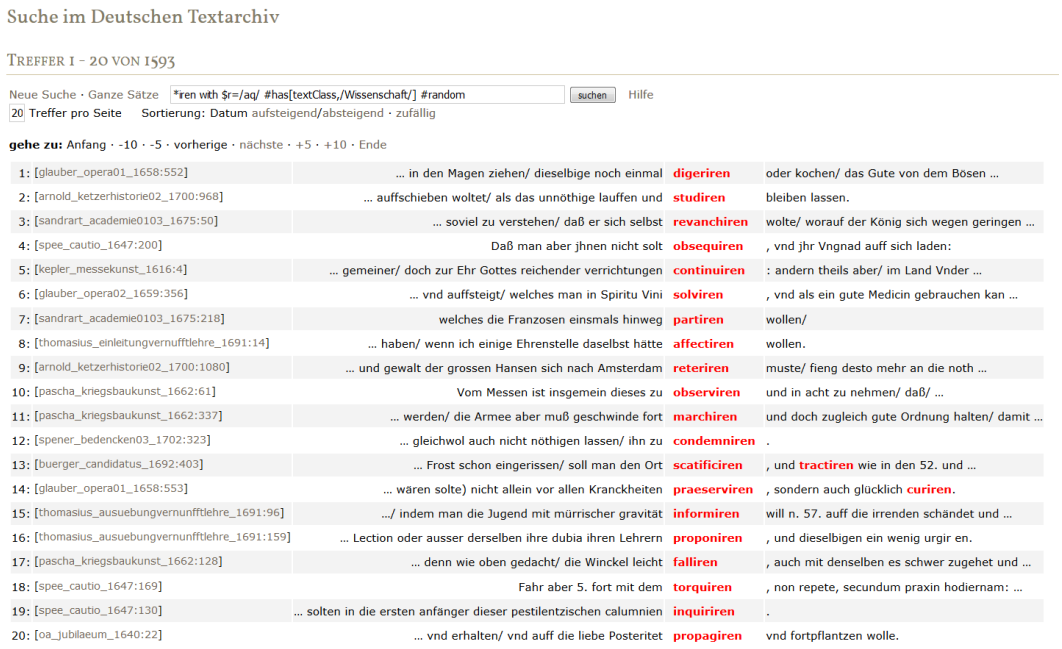

*Abb. 3: Buchstartseite im DTA mit ausführlichen Metadaten und dem Kommentar zur zugrunde gelegten Ausgabe, hier: Goethe 1790.*

Das DTABf ist mittlerweile als TEI-Auszeichnungsformat auch über das DTA hinaus bekannt. So stellt es beispielsweise das Best-Practice-Format für die Auszeichnung historischer Texte in CLARIN-D dar.[16]

Die strukturelle Auszeichnung der Volltexte ermöglicht z. B. deren Darstellung äquivalent zur Vorlage. Die Semantik der Auszeichnung kann aber auch bei Suchanfragen über das Korpus mit einbezogen werden. Beispielsweise können Suchen auf bestimmte Textsorten (z. B. Belletristik oder Gebrauchsliteratur), bestimmte Textbereiche (z. B. Überschriften, Bühnenanweisungen, Briefe, Anmerkungen) oder auch einzelne Subkorpora (z. B. auf das im Rahmen des Projekts AEDit[17] erstellte Korpus frühneuzeitlicher Leichenpredigten) beschränkt und miteinander verglichen werden.

16 CLARIN-D User Guide, http://www.clarin-d.de/de/hilfe/benutzerhandbuch, chapter 6, section 2 (Text Corpora).

17 AEDit Frühe Neuzeit: Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit, PURL: http://diglib.hab.de/?link=029. Das DTA ist einer der Kooperationspartner in diesem Projekt.



Suche im Deutschen Textarchiv

TREFFER 1 - 20 VON 1593

Neue Suche · Ganze Sätze *ren with $r=/aq/ #has[textClass,/Wissenschaft/] #random suchen Hilfe

20 Treffer pro Seite Sortierung: Datum aufsteigend/absteigend · zufällig

**gehe zu:** Anfang · -10 · -5 · vorherige · nächste · +5 · +10 · Ende

| | | | |
|---|---|---|---|
| 1: [glauber_opera01_1658:552] | ... in den Magen ziehen/ dieselbige noch einmal | digeriren | oder kochen/ das Gute von dem Bösen ... |
| 2: [arnold_ketzerhistorie02_1700:968] | ... auffschieben woltet/ als das unnöthige lauffen und | studiren | bleiben lassen. |
| 3: [sandrart_academie0103_1675:50] | ... soviel zu verstehen/ daß er sich selbst | revanchiren | wolte/ worauf der König sich wegen geringen ... |
| 4: [spee_cautio_1647:200] | Daß man aber jhnen nicht solt | obsequiren | , vnd jhr Vngnad auff sich laden: |
| 5: [kepler_messekunst_1616:4] | ... gemeiner/ doch zur Ehr Gottes reichender verrichtungen | continuiren | : andern theils aber/ im Land Vnder ... |
| 6: [glauber_opera02_1659:356] | ... vnd auffsteigt/ welches man in Spiritu Vini | solviren | , vnd als ein gute Medicin gebrauchen kan ... |
| 7: [sandrart_academie0103_1675:218] | welches die Franzosen einsmals hinweg | partiren | wollen/ |
| 8: [thomasius_einleitungvernunfftlehre_1691:14] | ... haben/ wenn ich einige Ehrenstelle daselbst hätte | affectiren | wollen. |
| 9: [arnold_ketzerhistorie02_1700:1080] | ... und gewalt der grossen Hansen sich nach Amsterdam | reteriren | muste/ fieng desto mehr an die noth ... |
| 10: [pascha_kriegsbaukunst_1662:61] | Vom Messen ist insgemein dieses zu | observiren | und in acht zu nehmen/ daß/ ... |
| 11: [pascha_kriegsbaukunst_1662:337] | ... werden/ die Armee aber muß geschwinde fort | marchiren | und doch zugleich gute Ordnung halten/ damit ... |
| 12: [spener_bedencken03_1702:323] | ... gleichwol auch nicht nöthigen lassen/ ihn zu | condemniren | . |
| 13: [buerger_candidatus_1692:403] | ... Frost schon eingerissen/ soll man den Ort | scatificiren | , und tractiren wie in den 52. und ... |
| 14: [glauber_opera01_1658:553] | ... wären solte) nicht allein vor allen Kranckheiten | praeserviren | , sondern auch glücklich curiren. |
| 15: [thomasius_ausuebungvernunfftlehre_1691:96] | .../ indem man die Jugend mit mürrischer gravität | informiren | will n. 57. auff die irrenden schändet und ... |
| 16: [thomasius_ausuebungvernunfftlehre_1691:159] | ... Lection oder ausser derselben ihre dubia ihren Lehrern | proponiren | , und dieselbigen ein wenig urgir en. |
| 17: [pascha_kriegsbaukunst_1662:128] | ... denn wie oben gedacht/ die Winckel leicht | falliren | , auch mit denselben es schwer zugehet und ... |
| 18: [spee_cautio_1647:169] | Fahr aber 5. fort mit dem | torquiren | , non repete, secundum praxin hodiernam: ... |
| 19: [spee_cautio_1647:130] | ... solten in die ersten anfänger dieser pestilentzischen calumnien | inquiriren | . |
| 20: [oa_jubilaeum_1640:22] | ... vnd erhalten/ vnd auff die liebe Posteritet | propagiren | vnd fortpflantzen wolle. |

*Abb. 4: Beispielhafte Suchanfrage: Suche nach Begriffen mit der Endung „iren", die durch eine Antiqua-Type von dem sie umgebenden Fraktur-Text abgegrenzt sind, beschränkt auf Texte der Kategorie „Wissenschaft"; Ergebnisse in KWIC-Ansicht mit zufälliger Sortierung.*

## III.4 AUTOMATISCHE LINGUISTISCHE ANALYSE

Anschließend an die strukturelle Auszeichnung der Texte im DTABf erfolgt die automatische linguistische Aufbereitung, welche die Tokenisierung, Lemmatisierung, die Bestimmung der Wortart (*Part-of-Speech, POS*[18]) und die automatisierte ‚Normierung' der historischen Schreibvarianten mittels der im DTA entwickelten Software CAB[19] umfasst. Die Annotation erfolgt mittels Standoff-XML-Markup. Die in dieser Weise vollzogene linguistische Aufbereitung ermöglicht nicht nur die Suche nach Wortformen ungeachtet ihrer möglicherweise vielfältigen (historischen) Schreibweisen, sondern erlaubt auch komplexe Suchanfragen, welche z. B. die Einbettung eines Suchterms in die umgebende Morphologie mit einbeziehen.

18 Das vom DTA für die linguistische Annotation verwendete Tagset basiert auf dem „Stuttgart/Tübinger Tagset" (STTS), siehe dazu *www.deutschestextarchiv.de/doku/pos*.

19 CAB: Cascaded Analysis Broker; vgl. Jurish 2012.


*Abb. 5: Beispielhafte Suchanfrage: Suche nach „grauen" und verwandten Formen, die bei der POS-Analyse als Verbform (oberes Bild) bzw. als Adjektiv (unteres Bild) klassifiziert wurden; Ergebnisse in KWIC-Ansicht mit zufälliger Sortierung.*

Die linguistische Analyse ermöglicht zudem die Visualisierung bestimmter Merkmale. Type- oder Lemmalisten und daraus erstellte ‚Wortwolken' (auch *word* oder *tag clouds* genannt) können hier einen leicht zugänglichen Überblick über das sprachliche Inventar eines Textes bzw. einer Gruppe von Texten liefern.


*Abb. 6: Lemmabasierte Wortwolke für Arnim 1806; angezeigt werden in diesem Beispiel nur Begriffe, die vom Part-of-Speech-Tagger als Substantiv (Klasse NN) klassifiziert wurden und eine Frequenz ≥ 3 haben. Die Schriftgröße der einzelnen Lemmata ist proportional zu deren Frequenz im Dokument.*

## III.5 WEBBASIERTE QUALITÄTSSICHERUNG (DTAQ)

Die Qualitätssicherung im DTA erfolgt zum einen im Vorhinein der Texterfassung durch ausführlich dokumentierte Richtlinien und die individuelle Vorbereitung aller Bilddateien für die Texterfassung, zum anderen auf vielfältige Weise summativ, d.h. im Anschluss an die Texterfassung (Geyken et al. 2012).

Den Kernpunkt der summativen Qualitätssicherung bildet die im DTA entwickelte webbasierte Qualitätssicherungsumgebung DTAQ.[20] Alle digitalisierten Volltexte des DTA werden zunächst in DTAQ integriert, wobei jedes Werk seitenweise zusammen mit der jeweiligen Faksimileausgabe in DTAQ präsentiert wird. Mehrere Repräsentationen der Volltexte können eingesehen werden: die HTML-Leseansicht, die TEI-XML-Ansicht, der mittels CAB automatisch normalisierte Text sowie eine Textfassung, welche die im Rahmen der linguistischen Analyse jeder Wortform zugeordneten Lemmata und POS-Tags anzeigt. In DTAQ ist es nun möglich, die verschiedenen Instanzen aller Texte, insbesondere aber die Transkription und Annotation aufgrund der Bildvorlage zu kontrollieren und Fehler mithilfe sogenannter Tickets zur Korrektur vorzuschlagen. Nutzern, die ihre eigenen Teilkorpora in DTAQ verwalten, kann außerdem der Zugriff auf inte-

[20] Für weitere Informationen zu DTAQ und zur Anmeldung siehe Wiegand 2014 sowie *http://deutschestextarchiv.de/dtaq/about*.

grierte Text- und XML-Editoren gewährt werden, sodass *ad hoc*-Korrekturen an den XML-Texten möglich sind. Auch die Kontrolle der ausführlichen Metadaten ist in DTAQ vorgesehen.

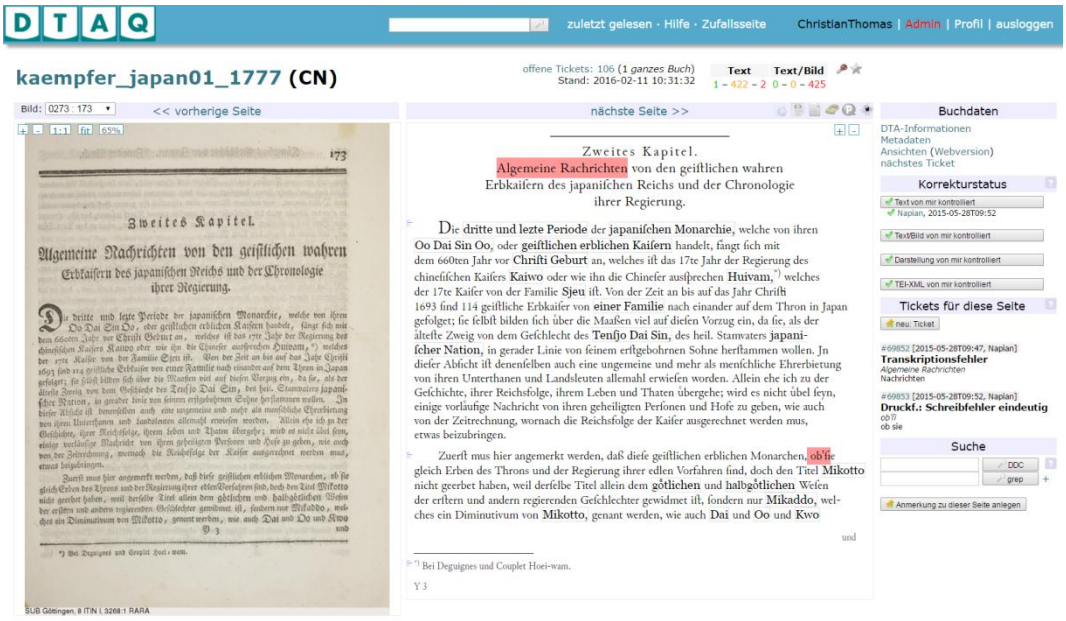



*Abb. 7: Ansicht einer Seite mit zwei offenen ‚Tickets' (einmal des Typs „Transkriptionsfehler", einmal des Typs „Druckfehler") aus Kaempfer 1777 in DTAQ.*

DTAQ ist nach Anmeldung allen Interessierten zugänglich.[21] Die Texte sind somit öffentlich verfügbar und können unter einer freien CC-Lizenz bereits aus DTAQ heraus heruntergeladen und nachgenutzt werden. Sie stellen dennoch eine Vorfassung der finalen Textausgaben dar, für die im Rahmen des DTA-Workflows noch weitere Korrekturgänge vorgesehen sind. Erst wenn diese Korrekturgänge abgeschlossen sind und die Werke eine verlässlich hohe Qualität aufweisen, werden sie auf der DTA-Webseite auch ohne Passwort zugänglich gemacht. Die passwortgeschützte ‚Vorveröffentlichung' der Texte in DTAQ ermöglicht zum einen die Einbeziehung der interessierten Community in die Qualitätssicherung, zum anderen einen kurzfristigeren freien Zugang zu den Werken, welche sich im DTA in Bearbeitung befinden.

Der letzte Schritt im Digitalisierungsworkflow des DTA ist die Veröffentlichung sämtlicher Texte auf der DTA-Webseite. Hier können sie eingesehen, einzeln oder als (Teil-)Korpora durchsucht und für die Nachnutzung heruntergeladen werden.

21 Zugang unter www.deutschestextarchiv.de/dtaq/.

# IV. ARBEITEN IM UND MIT DEM DTA

## IV.1 NUTZUNG DER DTA-INFRASTRUKTUR FÜR DIE VOLLTEXTDIGITALISIERUNG

*a) Das DTA-oXygen-Framework DTAoX.* Für die DTABf- und somit TEI-konforme Texterfassung durch Beitragende, die außerhalb des DTA arbeiten, stellt das DTA verschiedene Hilfsmittel bereit, welche die Arbeitsschritte zur Volltextdigitalisierung erleichtern und die korrekte Anwendung der DTA-Richtlinien sicherstellen sollen. Eines dieser Hilfsmittel ist das DTA-oXygen-Framework DTAoX.[22] Dabei handelt es sich um eine Anpassung für den Autormodus des oXygen-XML-Editors, die auf die Richtlinien des DTA-Basisformats abgestimmt ist und das Rendering der Textauszeichnungen erlaubt, sodass diese typographisch visualisiert werden können. Eine Werkzeugleiste, in welcher die DTABf-Elemente nach semantischen und formalen Kriterien sortiert sind, erleichtert den Einstieg in die Arbeit und unterstützt die Verwendung von DTABf-Elementen im Text. Ein Farbschema zeigt an, welche Auszeichnungstiefe durch Verwendung eines Elements erreicht wurde.[23]

[22] DTAoX ist zugänglich zum Download unter *www.deutschestextarchiv.de/doku/software#dtaox*.

[23] Die Elemente des DTABf sind in drei Auszeichnungslevel unterteilt: 1. obligatorische (z. B. `<body>`, `<div>`, `<p>`), 2. empfohlene (z. B. `<cit>`, `<hi>`, `<lb>`), 3. fakultative (z. B. `<choice>`, `<persName>`, `<foreign>`). Ein viertes Level kennzeichnet „unzulässige" TEI-Elemente, also solche, die explizit zugunsten anderer TEI-Lösungen aufgegeben wurden. Die Korpora des DTA sind konsequent bis zur Auszeichnungstiefe von Level 2 annotiert, während Level 3-Auszeichnungen im Projektkontext des DTA nicht durchgängig geleistet werden können. Dennoch werden Level 3-Elemente im DTABf angeboten, um Nutzern und Nutzerinnen des DTABf auch die tiefere Texterschließung zu ermöglichen. S. zu den DTABf-Levels auch die Übersicht unter: *www.deutschestextarchiv.de/doku/basisformat_table*.

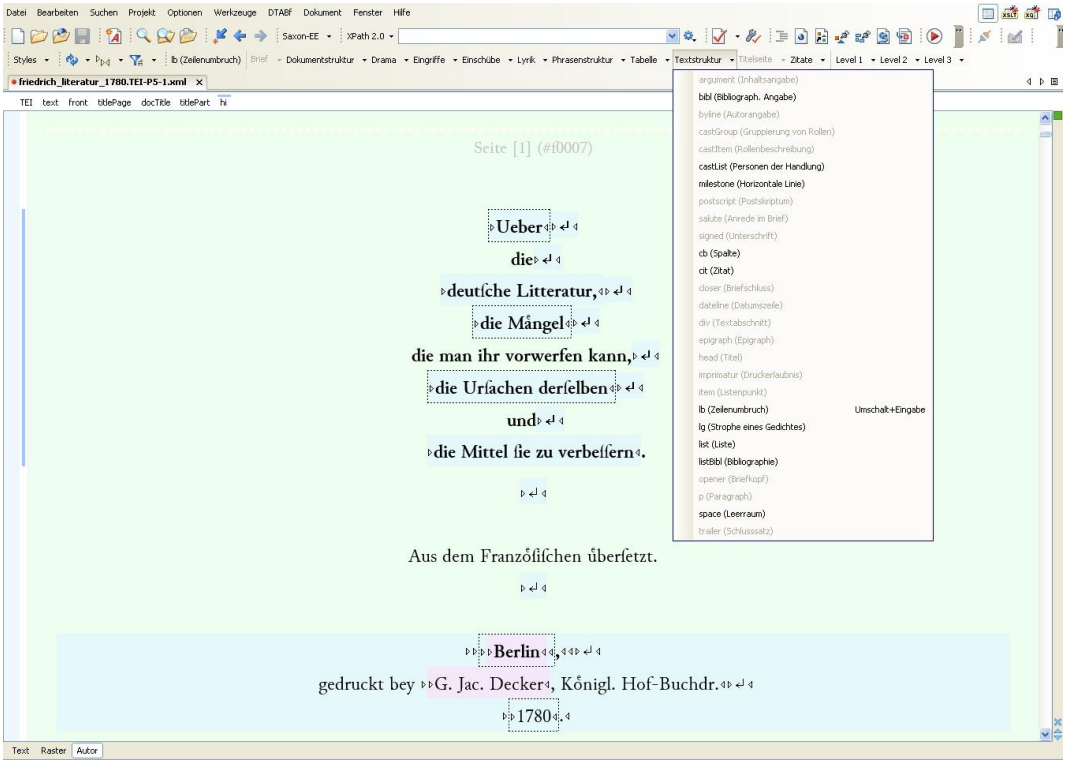



*Abb. 8: Ansicht des DTA-oXygen-Frameworks DTAoX: Titelseite aus Friedrich II. 1780.*

*b) Textkorrektur und tiefere Annotation in DTAQ.* Für die weitere Bearbeitung von DTABf-Dokumenten stehen in DTAQ zwei verschiedene Editoren bereit: Ein Texteditor ermöglicht die reine Transkription und Textkorrektur in der HTML-Ansicht des jeweiligen Textes. Dabei steht die Arbeit am Volltext im Vordergrund. Der Korrektur der XML-Annotationen dient ein weiterer integrierter XML-Editor. Darin ist es möglich, XML-Strukturen, welche die Seitenebene nicht überschreiten, anzubringen bzw. zu korrigieren. Über Buttons für die am häufigsten verwendeten Elemente können diese leicht in den Text eingefügt werden. In einem GIT-Repository werden alle Änderungen am Text gespeichert und können so nachvollzogen und redaktionell betreut werden. Es ist daher möglich, allen Nutzern von DTAQ die Textkorrektur und Transkription über die integrierten Editoren zu erlauben, da Fehler, die im Zuge der Korrekturschritte möglicherweise unterlaufen, nachvollzogen und wiederum behoben werden können.

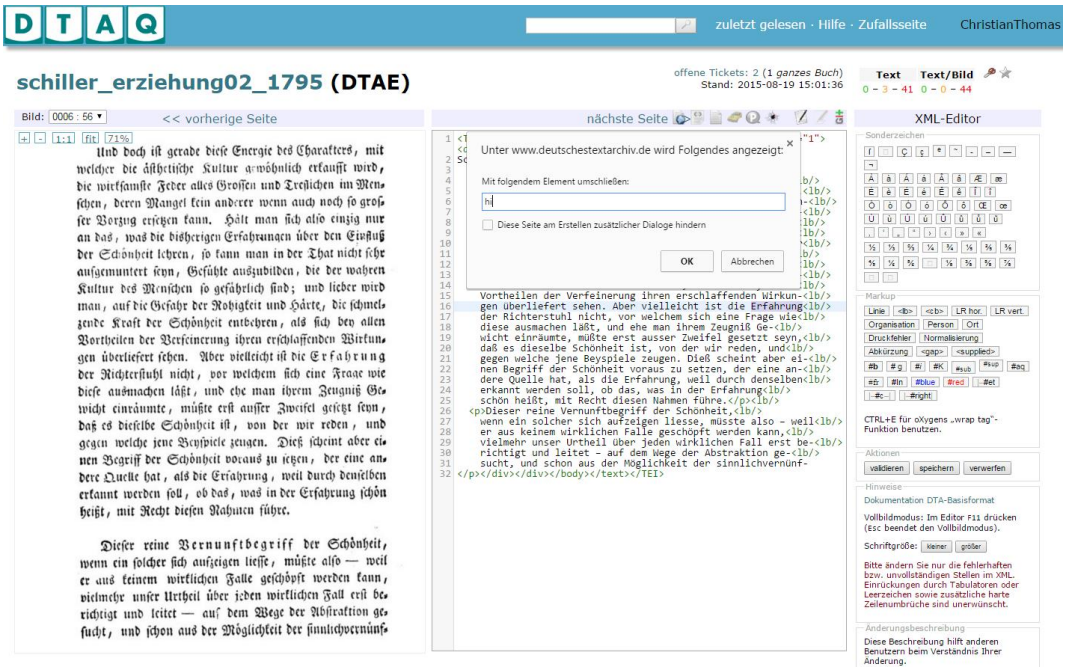



*Abb. 9: Nacherfassen einer Hervorhebung durch Sperrdruck in Schiller 1795 mithilfe des integrierten XML-Editors in DTAQ.*

*c) Formular zur Metadatenerfassung.* Zur Verbesserung der Recherchierbarkeit der Texte, der Auffindbarkeit der zugrundeliegenden Quellen sowie zur Dokumentation von Richtlinien, editorischen Entscheidungen etc. sieht das DTABf ausführliche Metadaten vor. Die DTABf-konforme Metadatenerfassung im TEI-Header wird jedoch daher schnell umfangreich und komplex. Um hier Fehler zu vermeiden und die Metadatenerfassung zu erleichtern, bietet das DTA ein webbasiertes Formular an, in welchem Metadaten zu verschiedenen Textsorten entsprechend vorgegebener Felder erfasst und im Anschluss an die Erfassung als TEI-Header ausgegeben und abgespeichert werden können.[24]

[24] Das DTA-Metadatenformular ist zugänglich unter: *www.deutschestextarchiv.de/dtae/submit/clarin*.

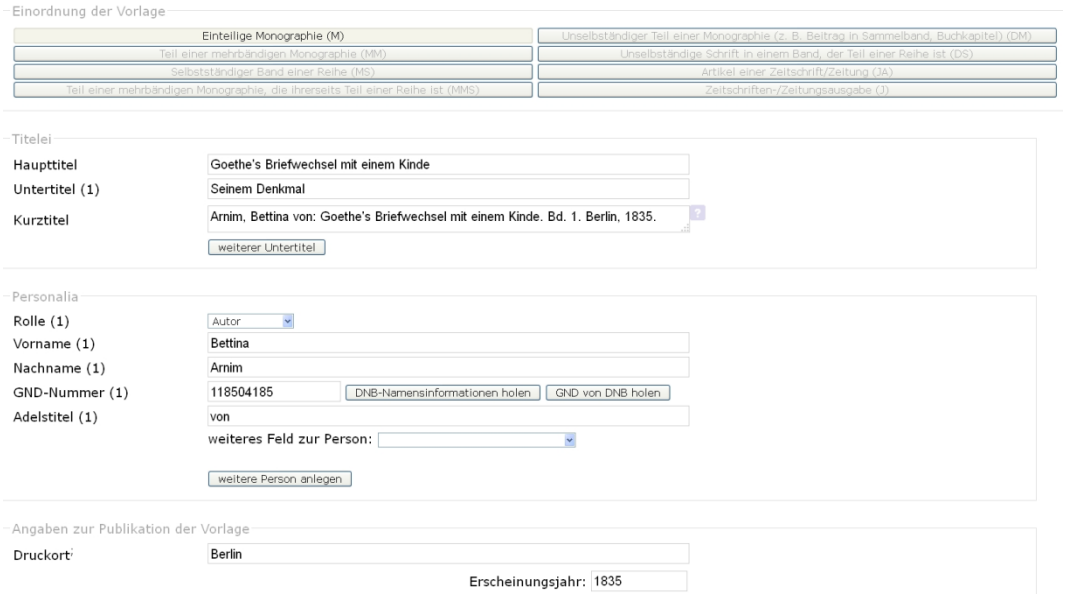



*Abb. 10: Webbasiertes Metadatenformular des DTA (http://www.deutschestextarchiv.de/dtae/submit/clarin)*

## IV.2 KORPUSRECHERCHE IM DTA

Um die Korpusrecherche im DTA zu ermöglichen, werden sämtliche Texte mit der Suchmaschine DDC (Dialing DWDS Concordancer) indiziert. DDC ermöglicht nicht allein einfache Wortsuchen, sondern auch komplexe Phrasensuchen, welche die linguistische Analyse und damit auch die orthographische Normierung mittels CAB sowie die TEI-Strukturierung einbeziehen.

So können zum Beispiel Termini in festgelegter oder variabler Schreibung, Wortarten oder GermaNet-Synsets[25] direkt nebeneinander, in einem bestimmten Abstand zueinander oder verknüpft durch Boolesche Operatoren als Phrasen aufgesucht werden. Auch die Position einer Phrase oder eines Terminus im Satz kann dabei festgelegt werden.

Darüber hinaus ist es möglich, inhaltliche Annotationen in die Suche einzubeziehen, sowohl durch die Einschränkung von Suchanfragen auf das Textmaterial in bestimmten TEI-Elementen als auch durch die Einbeziehung bestimmter Metadaten.

Die DDC-Syntax und die sich daraus ergebenden Möglichkeiten der Suche sind ausführlich dokumentiert.[26]

[25] GermaNet ist ein lexikalisch-semantisches Netz, das Substantive, Verben und Adjektive entsprechend der ihnen zugrundeliegenden Konzepte in Synsets gruppiert und diese zueinander in Relation stellt. URL: *http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml*; vgl. Hamp/Feldweg 1997, Henrich/Hinrichs 2010.

[26] *http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe*; vgl. auch *http://www.ddc-concordance.org/*.

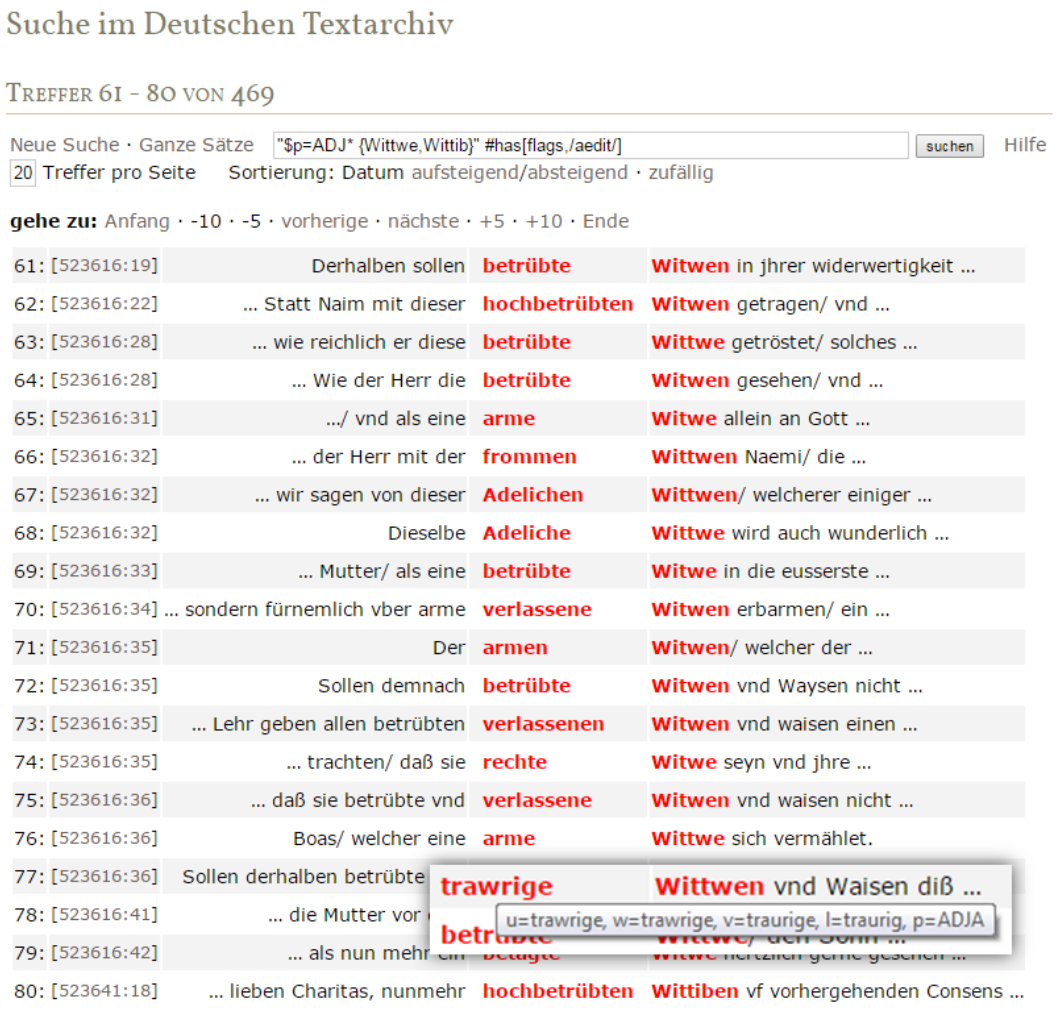



*Abb. 11: DDC-Suchanfrage: Suche nach einem beliebigen Adjektiv, gefolgt vom Terminus „Wittwe" bzw. „Wittib" im Subkorpus AEDit, das Leichenpredigten aus dem Bestand der ehemaligen Stadtbibliothek Breslau enthält*[27]

## V. LITERATURVERZEICHNIS

### 1. Primärquellen

Arnim 1806: Arnim, Achim von/Clemens Brentano: Des Knaben Wunderhorn. Bd. 1. Heidelberg 1806, in: Deutsches Textarchiv *<http://www.deutschestextarchiv.de/arnim_wunderhorn01_1806>*, abgerufen am 16.07.2014.

Friedrich II. 1780: Friedrich II., König von Preußen: Über die deutsche Literatur [..] Aus dem Französischen übersetzt (Übers. [Christian Konrad Wilhelm Dohm]). Berlin 1780, in: Deutsches Textarchiv *<http://www.deutschestextarchiv.de/friedrich_literatur_1780>*, abgerufen am 16.07.2014.

[27] Siehe oben, Anm. 14.

Gellert 1747: Gellert, Christian Fürchtegott: Das Leben der Schwedischen Gräfinn von G.***. 2 Bde. Leipzig 1747/48, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/gellert_leben01_1747*>, <*http://www.deutschestextarchiv.de/gellert_leben02_1748*>, beide abgerufen am 16.07.2014

Goethe 1774: Goethe, Johann Wolfgang von: Die Leiden des jungen Werthers. 2 Bde. Leipzig 1774, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/goethe_werther01_1774*>, <*http://www.deutschestextarchiv.de/goethe_werther02_1774*>, beide abgerufen am 16.07.2014.

Goethe 1790: Goethe, Johann Wolfgang von: Versuch die Metamorphose der Pflanzen zu erklären. Gotha 1790, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/goethe_metamorphose_1790*>, abgerufen am 16.07.2014.

Kaempfer 1777: Kaempfer, Engelbert: Geschichte und Beschreibung von Japan. Hrsg. v. Christian Wilhelm von Dohm. Bd. 1. Lemgo 1777, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/kaempfer_japan01_1777*>, abgerufen am 13.04.2016.

Karsch 1764: Karsch, Anna Luise: Auserlesene Gedichte. Berlin 1764, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/karsch_gedichte_1764*>, abgerufen am 16.07.2014.

Knigge 1788: Knigge, Adolph von: Ueber den Umgang mit Menschen. 2 Bde. Hannover 1788, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/knigge_umgang01_1788*>, <*http://www.deutschestextarchiv.de/knigge_umgang02_1788*>, abgerufen am 16.07.2014.

Liebig 1840: Liebig, Justus von: Die organische Chemie in ihrer Anwendung auf Agricultur und Physiologie. Braunschweig 1840, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/liebig_agricultur_1840*>, abgerufen am 16.07.2014.

Liebig 1842: Liebig, Justus von: Die organische Chemie in ihrer Anwendung auf Physiologie und Pathologie. Braunschweig 1842, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/liebig_physiologie_1842*>, abgerufen am 16.07.2014.

Röntgen 1896a: Röntgen, Wilhelm Conrad: Ueber eine neue Art von Strahlen. Vorläufige Mittheilung. 2. Aufl. Würzburg 1896, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/roentgen_strahlen_1896*>, abgerufen am 16.07.2014.

Röntgen 1896b: Röntgen, Wilhelm Conrad: Ueber eine neue Art von Strahlen. Fortsetzung, in: Sitzungsberichte der Würzburger Physik.- medic. Gesellschaft. Würzburg 1896, S. 1–9, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/roentgen_strahlen02_1896*>, abgerufen am 16.07.2014.

Röntgen 1897: Röntgen, Wilhelm Conrad: Weitere Beobachtungen über die Eigenschaften der X-Strahlen. In: Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin. Erster Halbband. Berlin 1897, S. 576–592, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/roentgen_weitere_1897*>, abgerufen am 16.07.2014.

Schiller 1784: Schiller, Friedrich: Kabale und Liebe. Mannheim 1784, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/schiller_kabale_1784*>, abgerufen am 16.07.2014.

Schiller 1795: Schiller, Friedrich von: Ueber die ästhetische Erziehung des Menschen in einer Reyhe von Briefen. Tl. 2, in: Schiller, Friedrich von (Hg.): Die Horen. Tübingen 1795, S. 51–94, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/schiller_erziehung02_1795*>, abgerufen am 16.07.2014.

Wedekind 1891: Wedekind, Frank: Frühlings Erwachen. Zürich 1891, in: Deutsches Textarchiv <*http://www.deutschestextarchiv.de/wedekind_erwachen_1891*>, abgerufen am 16.07.2014.

## 2. Sekundärliteratur

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

Haaf, Susanne/Frank Wiegand/Alexander Geyken: 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 (jTEI) 4 (2013), *http://jtei.revues.org/739*.

Geyken, Alexander/Susanne Haaf/Frank Wiegand: The DTA 'base format'. A TEI-Subset for the Compilation of Interoperable Corpora, in: 11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing, Proceedings of the Conference, hg. von Jeremy Jancsary, Wien 2012 (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5), *http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf*.

Hamp, Birgit/Helmut Feldweg: 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, Madrid, 1997.

Henrich, Verena/Erhard Hinrichs: GernEdiT - The GermaNet Editing Tool, in: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta, May 2010, pp. 2228-2235, (*http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf*)

Bryan Jurish: Finite-state Canonicalization Techniques for Historical German. Dissertation zur Erlangung des akademischen Grades doctor philosophiæ (Dr. phil.), Universität Potsdam, 2012, urn:nbn:de:kobv:517-opus-55789, *http://opus.kobv.de/ubp/volltexte/2012/5578/*.

TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.6.0, 20. Jan. 2014. *http://www.tei-c.org/Vault/P5/2.5.0/doc/tei-p5-doc/en/html/*.

Thomas, Christian/Frank Wiegand: 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*. (=Corpus Linguistics and Interdisciplinary Perspectives on Language 5) Tübingen 2015, S. 181–196.

Wiegand, Frank: Qualitätssicherung im Deutschen Textarchiv. Vortrag auf der 3. DGI-Konferenz, Frankfurt/M. 8. Mai 2014, *http://www.deutschestextarchiv.de/files/wiegand-dgi-2014-05-08.pdf*.