Alexander Geyken, 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

Abstract

This paper deals with the issue of quality assurance in very large, XML/TEI-encoded full-text collections. The text corpus edited by the DFG-funded project *Deutsches Textarchiv* (henceforth: DTA), a large and still growing reference corpus of historical German, is a fine example of such a collection. The following remarks focus on text prepared in a *Double-Keying*-process, since the major part of the DTA-corpus is compiled by applying this highly accurate method. An extensive and multi-tiered approach, which is currently applied by the DTA for the analysis and correction of errors in double-keyed text, is introduced. The process of quality assurance is pursued in a formative way in order to prevent as many errors as possible, as well as in a summative way in order to track errors which nevertheless may have occurred in the course of full-text digitization. To facilitate the latter, DTAQ, a web-based, collaborative tool for finding and commenting errors in the corpus, was developed. On the profound basis of practical experience in the past four years, the preliminaries and possible methods of conducting a widespread quality assurance are being discussed.

Einleitung

Das Projekt *Deutsches Textarchiv* (im Folgenden: DTA), gefördert von der Deutschen Forschungsgemeinschaft (DFG), ist Teil des Zentrums Sprache der *Berlin-Brandenburgischen Akademie der Wissenschaften* (BBAW).[1] Ziel des DTA ist es, einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus der Zeit von circa 1650–

[1] Zum Projekt vgl. die Homepage des Deutschen Textarchivs [1] sowie Geyken et al. (2011).

1900 nach den Erstausgaben im Volltext zu digitalisieren und somit den Grundstock für das erste umfassende historische Referenzkorpus des Deutschen bereitzustellen. Um den historischen Sprachstand möglichst genau abzubilden, werden als Vorlage für die Digitalisierung in der Regel die ersten selbständigen Ausgaben der jeweiligen Werke zugrunde gelegt. Die Volltexterfassung erfolgt möglichst vorlagengetreu und unter Verzicht auf textkritische Eingriffe und Kommentierungen. Hinsichtlich der Entstehungszeit der für das DTA erfassten Texte sowie in Bezug auf die dabei berücksichtigten Textsorten wird eine größtmögliche Ausgewogenheit angestrebt. Das DTA ergänzt als historische Komponente das ebenfalls an der BBAW erstellte, auf das 20./21. Jahrhundert fokussierte Kernkorpus des Digitalen Wörterbuchs der deutschen Sprache (DWDS).[2]

Mit dieser Zielsetzung unterscheidet sich das DTA von anderen, im Wesentlichen opportunistischen Textsammlungen wie Google Bücher, Wikisource, Zeno.org oder Gutenberg-DE. Zudem zeichnet sich das DTA durch die sorgfältige Auswahl der Digitalisierungsvorlagen, die Qualität der sie dokumentierenden Metadaten, die hohe Erfassungsgenauigkeit der präsentierten Texte sowie durch deren linguistische Erschließung gegenüber den genannten Textsammlungen aus. Die Kodierung der Volltexte erfolgt im DTA-Basisformat [2], das auf dem Schema der Text Encoding Initiative [3] (Version TEI-P5) basiert. Damit ist die Nachnutzbarkeit der Daten (beispielsweise als Grundlage editorischer Unternehmungen) und deren Austausch mit anderen Projekten sichergestellt.

Das Projekt DTA wird bis zum Ende seiner Laufzeit mehr als 1.300 Werke digitalisiert haben. Derzeit (Februar 2012) sind mehr als 700 Bände im Volltexte entsprechend dem DTA-Format vorhanden. Mit rund 400 Millionen Zeichen (71 Millionen Tokens; 1,65 Millionen Types) stellt das DTA bereits jetzt das größte historische XML/TEI-kodierte Volltextkorpus in deutscher Sprache dar. Die Texte des DTA sind über das Internet frei zugänglich und aufgrund ihrer Aufbereitung durch (computer-)linguistische Methoden schreibweisentolerant über den gesamten verfügbaren Bestand durchsuchbar.

Ein Großteil der im DTA zu digitalisierenden Drucke liegt in Frakturschrift vor. Aufgrund der heterogenen Zusammensetzung – das Korpus umfasst wissenschaftliche Texte unterschiedlicher Disziplinen ebenso

[2] Vgl. Geyken (2007).

wie Gebrauchstexte und Belletristik – muss das DTA mit einer Vielzahl verschiedener Strukturen umgehen. Daher erfolgte die Texterfassung im DTA bislang je nach Zustand der Vorlage mittels unterschiedlicher Verfahren.

Ein kleiner Teil des DTA-Korpus, bestehend aus einfach strukturierten Texten aus der Zeit von 1780 bis 1900, wurde mittels der OCR-Software ABBYY *FineReader* mit anschließender manueller Nachkorrektur erfasst. *FineReader* verfügt über ein Frakturmodul, sodass neben Antiqua-Texten auch gut lesbare Frakturtexte für das OCR-Verfahren verwendet werden konnten.

Der weitaus größere Teil der Vorlagen im DTA-Korpus enthält jedoch schwierige Strukturen und/oder ist in Typen gesetzt, die für eine OCR-Software nur schwer entzifferbar sind. Daher wurden etwa 75% der bislang im DTA verfügbaren Texte zeichenweise manuell im sogenannten *Double-Keying*-Verfahren erfasst.

Für das *Double-Keying*-Verfahren wird im Allgemeinen eine sehr hohe Erfassungsgenauigkeit angenommen.[3] Tatsächlich ergaben erste exemplarische Untersuchungen am DTA-Korpus, dass die Erfassungsgenauigkeit der mittels OCR erfassten und manuell korrigierten Texte deutlich geringer ist als bei Texten, die im *Double-Keying*-Verfahren erfasst wurden.[4] Abschätzungen von Erfassungsgenauigkeiten greifen jedoch in der Regel für eine umfassende Evaluation zu kurz, da sie sich lediglich auf die Zeichenkorrektheit, nicht jedoch auf das XML-Markup des Textes beziehen. Der folgende Beitrag ist daher der Frage nach typischen Fehlerquellen und -kategorien bei der Erfassung und Annotation historischer Texte unterschiedlicher Textsorten gewidmet. Dabei werden exemplarisch Verfahren der Fehlerermittlung, Möglichkeiten der Fehlerklassifikation und -vermeidung sowie Methoden der formativen (das heißt antizipierenden) und summativen (das heißt retrospektiven) Qualitätssicherung dargestellt. Hierfür wird der gesamte Prozess, von der Bilddigitalisierung über die Erfassung des Textes, dessen semiautomatische Aufbereitung und Annotation bis hin zur Publikation betrachtet und auf mögliche Fehlerquellen hin untersucht. Aufgrund der gesammelten Projekterfahrungen der vergangenen vier Jahre werden Strategien zur

[3] Vgl. auch die DFG Practical Guidelines on Digitization [6]: 11: »[Double Keying] allows transcription accuracies of up to 99.997%, i.e. virtually error-free texts.«

[4] Zur Problematik der geringen Erfassungsgenauigkeit beim OCR-Verfahren und der daraus folgenden Beeinträchtigung textbezogener Anwendungen vgl. z. B. Tanner et al. (2009); zu Methoden der Verbesserung von OCR-Ergebnissen vgl. Holley (2009).

effektiven Vermeidung oder Behebung verschiedener Typen von Fehlern beschrieben.

Qualitätssicherung im DTA

Die Qualitätssicherung ist ein Vorgang, der sämtliche Arbeitsschritte der Volltextdigitalisierung begleiten und somit nicht allein Möglichkeiten zur Ermittlung und Korrektur von entstandenen Fehlern, sondern auch Verfahren zu deren vorausschauender Vermeidung umfassen sollte. Die Arbeitsschritte der Volltextdigitalisierung im DTA lassen sich in vier Arbeitsbereiche einteilen:

1. Die Vorbereitung der Vorlagen für die Texterfassung und Annotation,
2. den eigentlichen Prozess der Texterfassung und Annotation,
3. die Nachbearbeitung der erfassten und annotierten Texte sowie deren Konvertierung in das DTA-Basisformat,
4. die anschließenden Korrekturgänge im Rahmen der Qualitätssicherungsumgebung DTAQ.

Diese einzelnen Arbeitsschritte sollen im Folgenden in Hinblick auf Möglichkeiten der Qualitätssicherung näher erläutert werden.

Vorbereitung der Vorlagen für die Texterfassung – formative Qualitätssicherung

Bereits bei der Vorbereitung der Digitalisate für die Texterfassung können Maßnahmen zur Fehlerprävention vorgenommen und somit eine dem Erfassungsprozess vorangestellte, formative Qualitätssicherung geleistet werden.

Die Maßnahmen zur formativen Qualitätssicherung umfassen im DTA die Kontrolle der Bilddateien, die den DTA-Volltexten zugrunde liegen, die DTA-Regelwerke mit den Richtlinien zur Texterfassung [4] und dem DTA-Basisformat [2] sowie die auf dieser Grundlage erstellten einzelfallbezogenen, bandspezifischen Anweisungen zur Erfassung und Auszeichnung.

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Bilddigitalisierung und -kontrolle

Die Bilddigitalisierung der in das Korpus aufzunehmenden Texte wird in der Regel von den Bibliotheken, welche die jeweilige Druckvorlage besitzen, vorgenommen. Ob ein Band ohne Textverlust digitalisiert werden kann oder ob die Vorlage gravierende Beschädigungen (zum Beispiel Schadstellen im Papier, Unvollständigkeit der Seiten et cetera) aufweist, wird vorab durch die Bibliotheken geprüft.

Die angefertigten Bilddigitalisate werden sodann im DTA einer Kontrolle unterzogen, bei der die Vollständigkeit, die korrekte inhaltliche und formale Abfolge sowie die Qualität der Scans überprüft werden. In einigen Fällen müssen die Reihenfolge der Scans korrigiert, einzelne Seiten nachgescannt oder defekte beziehungsweise fehlende Seiten aus einem gewissenhaft geprüften, druckidentischen Alternativexemplar ergänzt werden. Derartige Eingriffe werden in den Metadaten dokumentiert.

DTA-Richtlinien zur Texterfassung und DTA-Basisformat

Die Erstellung des elektronischen DTA-Volltextes setzt sich aus den miteinander kombinierten Arbeitsgängen der Transkription und Annotation zusammen. Die dabei zugrunde liegenden Richtlinien sind ausführlich dokumentiert und mit zahlreichen Beispielen versehen.

Die Vorschriften zur Transkription der Texte sind in den *DTA-Richtlinien zur Texterfassung* festgelegt. Diesen Richtlinien liegt das sprachwissenschaftlich motivierte Interesse zugrunde, den historischen Sprachstand der Texte möglichst verlustfrei zu erfassen. Sie folgen daher dem Prinzip größtmöglicher Bewahrung des Vorlagentextes unter Verzicht auf Normalisierungen bei gleichzeitiger Konzentration auf lexikalische Gegebenheiten. Die Zahl der (unvermeidbaren) Interpretationen typographischer Gegebenheiten soll dabei so gering wie möglich gehalten werden.

Die XML-Kodierung der Texte erfolgt im sogenannten *DTA-Basisformat*, einem XML-Schema auf der Grundlage der P5-Richtlinien der TEI[5]. Die Empfehlungen der TEI werden entsprechend den spezifischen Annotationsbedürfnissen des DTA präzisiert. So enthält das *DTA-Basisformat* für spezifische textuelle Phänomene möglichst eindeutige, die vorhandenen Interpretationsspielräume weitgehend einschrän-

kende Kodierungsvorschriften, um die Einheitlichkeit bei der Auszeichnung ähnlicher textueller Phänomene innerhalb des DTA-Korpus sicherzustellen.[5]

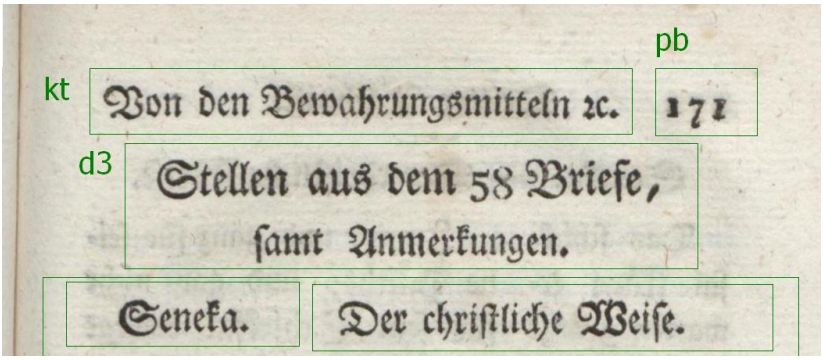
Hinsichtlich der Auszeichnung struktureller und typographischer Informationen der Vorlage folgen die Annotationsrichtlinien des DTA-Basisformats dem Prinzip »Struktur vor Typographie«, das heißt. es gilt zunächst, die Strukturierung der Vorlage mittels semantisch-strukturierender Auszeichnungen abzubilden. Lassen sich Strukturierungen nicht semantisch interpretieren, wird von diesem Prinzip abgewichen, und die typographischen Gegebenheiten der Vorlage werden abgebildet. So kann zum Beispiel eine Einrückung auf ein Zitat, aber auch einen Vers oder eine Widmung kennzeichnen, und wird entsprechend ausgezeichnet. Kann im Erfassungsprozess die Semantik jedoch nicht identifiziert werden, wird die betreffende Passage lediglich als »eingerückt« markiert. Ferner werden Überschriften in Bezug auf ihre hierarchische Stellung markiert, während typographische Hervorhebungen wie Schriftgröße, Zentrierung et cetera unberücksichtigt bleiben.[6]

Um den Erfassern die XML-Auszeichnung der Texte zu erleichtern und Fehlern vorzubeugen, werden die Digitalisate im Anschluss an die Scankontrolle mit einer im DTA entwickelten Software bearbeitet, die es ermöglicht, Textpassagen zu umrahmen und mit Beschriftungen zu versehen.

[5] Diese Spezifikation des DTA-Basisformats trägt dem Problem fehlender Interoperabilität von XML/TEI-kodierten Texten Rechnung; vgl. Unsworth (2011).

[6] Der Ansatz »Struktur vor Typographie« spielt auch für die spätere Darstellung der Texte im HTML-Format eine Rolle. Zum Beispiel werden alle Überschriften einheitlich entsprechend moderner typographischer Standards formatiert, während die evtl. in die Annotation übernommenen typographischen Besonderheiten (zentriert, fett, gesperrt, …) für die Darstellung ignoriert werden.

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)



Screenshot: Vorlage mit hervorgehobenen Strukturbereichen (Zonen) Bildausschnitt aus Sailer, Johann Michael: Ueber den Selbstmord. Für Menschen, die nicht fühlen den Werth, ein Mensch zu sein. München: Lentner, 1785, S. 171. [14]

Mit dem *Zoning Tool* (ZOT) ist es möglich, diejenigen Strukturbereiche hervorzuheben, die bei der Texterfassung durch Annotationen vom »reinen« Text abgesetzt werden sollen. Die Annotationen folgen dabei einem reduzierten Markup, dem das DTA-Basisformat zugrunde liegt.

Zusätzlich zu den genannten Regelwerken wird für jeden Text ein Begleitblatt erstellt, in welchem auf strukturelle und typographische Eigenheiten des jeweiligen Bandes hingewiesen wird. Dabei werden unter anderem Hinweise zur Wiedergabe der Initialen in der spezifischen Drucktype gegeben, Sonderzeichen geklärt, typographische Besonderheiten (zum Beispiel der Wechsel zu einer anderen Frakturtype zum Zweck der Hervorhebung) exemplarisch dargestellt oder auf problematische Gliederungen des Textes hingewiesen. Dies stellt einerseits sicher, dass strukturrelevante Abschnitte als solche erkannt werden, und gewährleistet andererseits die konsequente Anwendung des Tagsets.

Diese Maßnahmen der formativen Qualitätssicherung, das heißt die Richtlinien zur Texterfassung, deren Konkretisierung in Begleitblättern, die Spezifizierung des DTA-*Basisformats* und die Hinweise auf dessen Anwendung mittels entsprechender Zonen, sollen Erfassungsfehlern vorbeugen und die Kohärenz der verwendeten Annotationen innerhalb des Korpus‘ gewährleisten. Dennoch bleibt die Texterfassung und -strukturierung aufgrund unterschiedlicher Faktoren, die im Folgenden erläutert werden, fehleranfällig.

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Texterfassung und Annotation – Fehlerquellen und Fehlertypen

Im Anschluss an die Vorbereitung der Vorlagen erfolgt die Texterfassung und Annotation im *Double-Keying*-Verfahren. Dieses Verfahren stellt, wie bereits erwähnt, für gedruckte Texte eine bewährte und zuverlässige Methode der Texterfassung dar. Bei der zweifachen Texterfassung durch voneinander unabhängig arbeitende Erfasser mit anschließendem teilautomatischem Abgleich beider Textfassungen können Flüchtigkeitsfehler weitgehend vermieden werden. Dennoch existieren typische Fehlerquellen, die entweder so grundsätzlicher Natur sind, dass sie auch beim *Double-Keying*-Verfahren innerhalb des Erfassungsprozesses eine Rolle spielen, oder die im Rahmen der weiteren Bearbeitungsschritte der Volltexte auftreten können. Für die Entwicklung von Verfahren zur Qualitätssicherung ist die Kenntnis möglicher Fehlerquellen und Fehlertypen erforderlich. Daher wurde im DTA zunächst auf der Grundlage manueller Korrekturgänge an einer exemplarischen Textmenge eine Fehlerkategorisierung erstellt.

Genauigkeit

Generell lassen sich anhand des DTA-Korpus‘ zwei Arten der Genauigkeit unterscheiden: die der Texterfassung (Zeichengenauigkeit) und die der Annotation (Markupgenauigkeit). Die Zeichengenauigkeit betrifft den Grad der Übereinstimmung des transkribierten Textes mit der Vorlage. Die Markupgenauigkeit spezifiziert den Grad der Übereinstimmung der formalen und inhaltlichen Annotationen eines Volltextes mit den Gegebenheiten der Vorlage einerseits und mit den DTA-Annotationsrichtlinien sowie dem zugehörigen XML-Schema andererseits.

Während zur Zeichengenauigkeit bei der Texttranskription im *Double-Keying*-Verfahren konkrete Werte angegeben werden, lässt sich die Genauigkeit des Markup bislang nicht in vergleichbarer Weise messen.[7] Hinzu kommt, dass in beiden Fällen eine eindimensionale Beurteilung der Genauigkeit kaum aussagekräftig ist, da für beide Bereiche verschie-

[7] Siehe oben, Anm. 5.

dene Fehlerkategorien existieren, die unterschiedlich gewertet werden müssen.

Zeichengenauigkeit

Auf der Ebene der Zeichengenauigkeit können zwei Fehlerkategorien unterschieden werden: Transkriptionsfehler und Druckfehler. Wie häufig Fehler dieser Kategorien auftreten, ist innerhalb von *Double-Keying*-Texten nochmals von äußeren Faktoren abhängig, etwa von der Kenntnis der Sprache des Textes: Des Deutschen unkundige Erfasser (Nichtmuttersprachler) sehen sich bei der Transkription vor andere Probleme gestellt als mit der deutschen Sprache vertraute Erfasser (Muttersprachler). Die Qualität der Transkriptionen kann zudem durch Erfahrungen in der Erfassung historischer deutscher Texte sowie durch sprachhistorische Kenntnisse beeinflusst werden.[8]

Transkriptionsfehler

Die Ähnlichkeit historischen Wortmaterials zur Gegenwartssprache bei gleichzeitig (leicht) abweichender Schreibung (zum Beispiel *wolte* vs. *wollte*; *vnd* vs. *und*) kann bei (ungeschulten) Muttersprachlern zur versehentlichen Setzung moderner Schreibweisen führen. Nichtmuttersprachler sind hingegen darauf angewiesen, die unbekannten Worte kontextfrei zeichenweise zu transkribieren, wodurch die Gefahr versehentlicher Modernisierungen reduziert wird. Allerdings kann es hierbei auf der Ebene der Zeichenerkennung zu Problemen kommen. Verschiedene Eigenschaften der Druckvorlage können dabei die Zeichenerkennung erschweren und zu Erfassungsfehlern führen:

- Beschädigungen der Vorlage (zum Beispiel Stockflecken, dünnes Papier, enge Bindung),

[8] Im DTA wird daher zum Zweck der Qualitätssicherung die Verzahnung verschiedener Kompetenzen angestrebt (sprachunkundige Erfasser, die zeichengenau transkribieren; sprachkundige, z. T. philologisch ausgebildete Korrekturleser).

- Ausführung des Digitalisats (zum Beispiel unzureichende Auflösung, Informationsverlust durch Bitonalität der Bildaufnahmen, Verzerrungen),
- Beschaffenheit des Druckbildes (zum Beispiel sehr kleiner Druck, unscharfe Zeichengrenzen).

Auch Besonderheiten auf typographischer Ebene können die Zeichenkennung beeinträchtigen:

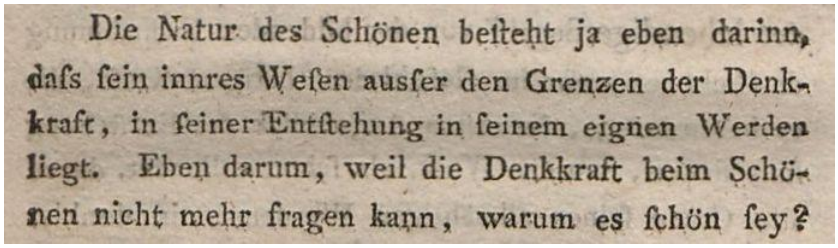
- Besonderheiten der Schrifttype
  - Verzierungen, insbesondere in der Frakturschrift (zum Beispiel bei Initialen),
  - Ähnlichkeit verschiedener Zeichen in der Frakturschrift (*C* vs. *E*, *M* vs. *W*, *R* vs. *K*, *V* vs. *B*, *e* vs. *c*, *k* vs. *t*, *v* vs. *p*). – Die Schwierigkeit, Zeichen einer Frakturtype korrekt zu erkennen, wird durch die Heterogenität des Textmaterials im DTA und die Herkunft der Drucke aus verschiedenen Offizinen verstärkt, da der daraus resultierende stete Wechsel der Frakturtypen einen Gewöhnungseffekt unmöglich macht.
  - Ähnlichkeit verschiedener Zeichen in Antiquaschrift (*e* vs. *o*, *I* vs. *l* beziehungsweise *1*, *0* vs. *O*).
- unscharfe Wortgrenzen: Wortzwischenräume können aufgrund der variablen Zeichenmenge je Zeile besonders in älteren Drucken hinsichtlich ihrer Breite variieren und sind daher nicht immer als solche zu erkennen. Dieses Problem stellt sich bei historischem Wortmaterial selbstverständlich auch für Muttersprachler (*hinzu thut* vs. *hinzuthut*, *dem felben* vs. *demfelben*).

Weitere Transkriptionsfehler ergeben sich aufgrund von Fehlinterpretationen der DTA-Richtlinien. So kann der Drucktext mit nachträglichen Ergänzungen wie handschriftlichen Bemerkungen oder eingeklebten Texten angereichert worden sein, die entgegen den DTA-Richtlinien fälschlich mit transkribiert wurden.

Druckfehler

Weiterhin können Fehler als Druckfehler bereits in der Vorlage tradiert worden sein. Die textkritische Leistung, diese Druckfehler aufzufinden und gegebenenfalls zu verbessern, kann durch die Erfasser nicht erbracht werden. Hier sind häufig nicht allein sprachliche, sondern zudem sprachhistorische (und andere philologische) Kenntnisse vonnöten, die

auch in die Bereiche der historischen Grammatik, Semantik und Orthographie reichen, wie im folgenden Beispiel.


Die Natur des Schönen beſteht ja eben darinn, daſs ſein innres Weſen ausſer den Grenzen der Denkkraft, in ſeiner Entſtehung in ſeinem eignen Werden liegt. Eben darum, weil die Denkkraft beim Schönen nicht mehr fragen kann, warum es ſchön ſey?

Karl Philipp Moritz: Über die bildende Nachahmung des Schönen. Braunschweig: Schul-Buchhandlung, 1788, S. 26.[15]

> Die Natur des Schönen beſteht ja eben darinn,
> daſs ſein innres Weſen ausſer den Grenzen der Denk¬
> kraft, in ſeiner Entſtehung in ſeinem eignen Werden
> liegt. […]

Die Schreibung *ſſ* neben *ſs* für das Phonem *ss* (heutige Schreibung *ss* beziehungsweise *ß*) könnte als Druckfehler missdeutet werden, ist jedoch historisch belegbar.[9]

Markupgenauigkeit

Im Zuge der Volltexterfassung werden die Texte des DTA-Korpus mittels eines reduzierten Markup, dem das DTA-Basisformat zugrunde liegt, strukturiert. Das reduzierte Markup, das aus gegenüber dem TEI-Standard stark verkürzten Tags und einigen Pseudo-Entitäten besteht, soll die Möglichkeit formaler Fehler beim Tagging reduzieren. Im Anschluss an die Volltexterfassung und -strukturierung werden die Texte über mehrere (größtenteils automatisierte) Zwischenschritte in DTA-Basisformat überführt. Dabei wird eine automatische Prüfung der XML-Validität durchgeführt, wodurch formale Fehler bei der XML-

[9] Vgl. Reichmann/Wegera (1993: 113, § L 53).

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Auszeichnung wie etwa fehlende schließende Tags oder Schreibfehler bei den Elementnamen behoben werden können.

Neben formalen Mängeln können jedoch verschiedene Fehler bei der strukturellen Auszeichnung auftreten:

1. Annotationsfehler als Transkriptionsfehler,
2. Setzung oder Nicht-Setzung von Strukturmerkmalen entgegen den DTA-Richtlinien beziehungsweise fehlerhafte Beurteilung struktureller Gegebenheiten der Vorlage,
3. Fehler bei der automatischen Bearbeitung und Umwandlung der XML-Dateien in das DTA-Basisformat.

Annotationsfehler als Transkriptionsfehler

Ein Teil der XML-Annotation ist eng mit der Texttranskription verknüpft. Dabei kann es zur fehlerhaften Abnahme typographischer Besonderheiten aus der Vorlage kommen (zum Beispiel nicht erkannter Fett- oder Kursivdruck, die fälschliche Auszeichnung einer Majuskel als Initiale, nicht erkannter Frakturwechsel).

Fehlerhafte Beurteilung struktureller Gegebenheiten der Vorlage

Strukturmerkmale werden bisweilen entgegen den DTA-Richtlinien nicht ausgezeichnet oder falsch interpretiert (zum Beispiel Bogensignaturen oder Kustoden werden nicht als solche erkannt; fälschliche Schließung von Paragraphen am Seitenende; Auszeichnung von Tabellenspalten als Spaltendruck; inkorrekte Setzung der Überschriftenebenen). Derartige Fehler können auch auf Fehlinterpretationen der DTA-Richtlinien zurückzuführen sein (zum Beispiel Auszeichnung von verzierten Trennlinien als Abbildungen, von gereimter Rede eines Sprechers im Drama als Gedicht; Dokumentation von Frakturwechsel in Überschriften und auf Titelblättern entgegen den DTA-Richtlinien).

Fehlerhafte Beurteilungen struktureller Gegebenheiten des Textes können zu Textverlust führen (zum Beispiel unerkannte Bildunterschriften).

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Umwandlung der XML-Dateien in das DTA-Basisformat

Das reduzierte Markup, mit dem die Texte bei ihrer Erfassung versehen werden, stellt eine starke Vereinfachung des DTA-Basisformats dar. Die Überführung in das DTA-Basisformat erfolgt halbautomatisch.

Beispiel: Im reduzierten Markup ist es möglich, Titel/Überschriften mittels <d>-Elementen auszuzeichnen:

```
<d1>[Titel Ebene 1]</d1>
<p>[Text Ebene 1]</p>
<d2>[Titel Ebene 2]</d2>
<p>[Text Ebene 2]</p>
<d2>[Titel Ebene 2]</d2>
<p>[...]</p>
<d1>[...]</d1>
```

Diese Struktur wird bei der Konvertierung nach XML/TEI in die folgende Struktur umgewandelt, wobei jedes -Element vor Beginn eines neuen -Elements höherer Ebene geschlossen wird:

```
<div n="1">
        <head>[Titel Ebene 1]</head>
        <p>[Text Ebene 1]</p>
        <div n="2">
                <head>[Titel Ebene 2]</head>
                <p>[Text Ebene 2]</p>
        </div>
        <div n="2">[...]</div>
</div>
<div n="1">[...]</div>
```

Bei der automatisierten Überführung der transkribierten Volltexte in das DTA-Basisformat kann es im Falle struktureller Besonderheiten, die einzelfallbezogen und daher in den angewandten Skripten noch nicht berücksichtigt worden sind, zu nachträglichen Annotationsfehlern kommen.

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Sonstiges

Fehler im Workflow: Im Anschluss an die Textauszeichnung mittels eines reduzierten Markups werden die DTA-Texte nicht allein durch verschiedene Konvertierungsschritte in das DTA-Basisformat überführt, sondern sie durchlaufen auch einige Nachbearbeitungsschritte, etwa die Bearbeitung beziehungsweise gegebenenfalls Nacherfassung von Zeichen, die durch die Erfasser als unleserlich markiert wurden, oder die Ersetzung von Pseudo-Entitäten durch die zugehörigen Unicode-Werte. Fehler im Workflow liegen vor, wenn einer dieser Nachbearbeitungsschritte nicht oder nicht korrekt ausgeführt worden ist.

Darstellungsfehler: Fehleranfällig kann schließlich auch die Darstellung der XML-Texte in einer HTML-Ansicht sein. So bietet das DTA eine parametrisierbare Leseansicht an, welche automatisiert Normalisierungen der historischen Schreibweisen vornimmt.[10] Diese Normalisierungen werden für oberflächlich gleichartige Fälle in gleicher Weise vorgenommen, was in Einzelfällen zu Darstellungsfehlern führen kann. So wird etwa die häufige Zeichenkombination *ſs* in das heute gebräuchliche *ß* überführt (zum Beispiel *groſs* → *groß*). In den (selteneren) Fällen, dass *ss* zu ergänzen wäre, kommt es folglich zu Darstellungsfehlern (zum Beispiel *laſsen* → *laßen*).

Qualitätssicherung innerhalb des Publikationsprozesses

Alle DTA-Texte durchlaufen den beschriebenen, standardisierten und ausführlich dokumentierten Prozess von der Auswahl der Digitalisierungsvorlage bis zum XML/TEI-strukturierten Volltext. Etliche Fehlerquellen auf diesem Weg wurden identifiziert und möglichst umgangen. Weitere, häufig auftretende Probleme können noch im Laufe des Publikationsprozesses, wie im Folgenden beschrieben, systematisch behoben werden.

---

[10] Parametrisierbare Leseansicht im DTA [7]. Das Erscheinungsbild der Texte liegt hier in der Hand der Nutzerinnen und Nutzer des DTA: So können einzelne Elemente von der Anzeige ausgenommen (z. B. der Zeilenfall) oder historische typografische Konventionen in der Anzeige verändert werden (*s* statt *ſ*, *r* statt rundem r (ꝛ), *ä*, *ö*, *ü* statt *a͛*, *o͛*, *u͛*).

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Verbesserung der Zeichengenauigkeit

Potentielle Schwierigkeiten im Bereich der Typographie und Vorlagenqualität sowie daraus resultierende Fehlerquellen können, wie gezeigt wurde, im Vorhinein durch Autopsie der Digitalisate erkannt und im jeweiligen Begleitblatt geklärt werden. Im Anschluss an die Texterfassung können Fehler durch die Suche nach typischen Zeichenketten (zum Beispiel *uud* vs. *und*) halbautomatisch ermittelt werden. Parallel zur Konvertierung nach TEI-P5 werden die Texte durch computerlinguistische Werkzeuge erschlossen. Auch hierdurch lassen sich bestimmte Arten von Fehlern identifizieren, beispielsweise durch Analyse der Zeichenketten, die im Lemmatisierungsprozess keine Zuordnung erhalten haben.

Wie oben beschrieben, werden Beschädigungen mit Textverlust und sonstige im Rahmen der Transkription ungelöste Fälle – zumindest soweit sie von den Erfassern als unleserlich markiert wurden – manuell überprüft. Sie können in den meisten Fällen anhand der Vorlage oder kontextbezogen ergänzt werden. Bisweilen ist hier die Konsultation einer weiteren, gleichartigen Textfassung vonnöten.

Da die Texterfassung durch sprachunkundige Erfasser erfolgt, bleiben viele für den Sprachkundigen offensichtliche Schreib-, Satz- und Druckfehler der Vorlage unerkannt: Sie werden ebenso exakt wie der übrige Text und insofern korrekt aus der Vorlage übernommen. Die in Druckfehlerverzeichnissen gesammelten Errata können systematisch behandelt werden. Alle übrigen Druckfehler können in der Regel ebenso wie die sonstigen Fehler auf der Ebene der Zeichengenauigkeit allein durch nachträgliches Kollationieren der digitalisierten Druckvorlage mit dem transkribierten Volltext ermittelt werden.

Verbesserung der Markupgenauigkeit

Die Voraussetzung für das Markup ist die Analyse des Textes hinsichtlich formal und inhaltlich strukturierender Merkmale. Diese Merkmale können auf der Grundlage des DTA-Basisformats im Zuge der Textvorbereitung gekennzeichnet werden.[11] Für einfache Strukturmerkmale genügt es in der Regel, eine exemplarische Vorstrukturierung vorzuneh-

[11] Siehe oben, Abschnitt *Vorbereitung der Vorlagen für die Texterfassung.*

men, die dann durch die Erfasser auf gleichwertige Fälle zu übertragen ist. Fehler, die bereits im Rahmen der Vorstrukturierung oder durch die selbständige Strukturierung durch die Erfasser auftreten, werden, sofern sie Auswirkungen auf der Darstellungsebene haben, im Zuge der nachträglichen Konvertierungsschritte in das DTA-Basisformat und in die HTML-Leseansicht bemerkt. Weniger offensichtliche Strukturierungsfehler können nur im Zuge der Nachkorrektur durch Abgleich des XML-Dokuments mit der Vorlage aufgefunden werden.

Verteiltes Korrekturlesen mit DTAQ – summative Qualitätssicherung

Das gründliche Korrekturlesen der Texte vor deren Veröffentlichung ist aus den von der DFG bewilligten Projektmitteln nicht zu leisten. Aus diesem Grund wurde im DTA zur Ergänzung der formativen Qualitätssicherung eine verteilte Korrekturumgebung (DTAQ)[8] entwickelt, mittels derer im Anschluss an die Veröffentlichung der Texte im DTA eine summative Qualitätssicherung vorgenommen werden kann.

Die webbasierte Arbeitsumgebung DTAQ soll das verteilte Korrekturlesen erleichtern und vorgenommene Anmerkungen in einem einheitlichen Kategorienschema speichern und verwalten. In DTAQ können die verschiedenen, oben beschriebenen Fehlertypen in Form von Tickets (ähnlich einem Bug-Tracker aus dem Bereich der Softwareentwicklung) gemeldet werden.

In DTAQ werden die Bilddigitalisate der bis dato bearbeiteten DTA-Texte zusammen mit den zugehörigen Volltexten angezeigt und können so seitenweise miteinander verglichen werden. Die Größe von Bildausschnitten und deren Position sind flexibel einstellbar; auch die Breite des Textfeldes kann der Nutzer seinen Wünschen entsprechend verändern. Der Text der gewählten Seite kann wahlweise als XML/TEI oder in der daraus generierten HTML-Ansicht angezeigt werden. Darüber hinaus bietet eine dritte Ansicht »CAB«[12] die Möglichkeit, die einzelnen Token der historischen Texte in automatisch erzeugter normalisierter Wortform (das heißt in heutiger Schreibung) darzustellen. Auch in der CAB-Ansicht sind Fehlermeldungen möglich und dienen der Verbesserung des Index. Weiterhin bietet DTAQ den Benutzern Part-of-Speech-

[12] Cascaded Analysis Broker, siehe dazu Jurish (2008, 2010, 2012).

Angaben und kann für selektierten Text entsprechende XPath-Werte generieren.

Die Entscheidung, was als Fehler anzusehen ist und was in Übereinstimmung mit den Vorgaben des DTA eventuell abweichend von der Vorlage dargestellt wurde, wird durch eine eigens zu diesem Zweck erstellte »Korrekturfibel«[9] unterstützt. Hierin werden die wichtigsten verwendeten XML-Tags und deren Erscheinungsbild in der HTML-Ansicht erklärt. Ferner werden Empfehlungen zu möglichen Arten des Korrekturlesens gegeben. Beispielsweise kann es neben der klassischen Text-Bild-Korrektur auch sinnvoll sein, sich vor allem auf Darstellungsprobleme oder auf häufig auftretende Fehler im Tagging zu konzentrieren.

Gemeldete Fehler können einer Kategorie zugeordnet, mit einer höheren oder niedrigeren Priorität versehen sowie einem bestimmten Bearbeiter zugewiesen werden. Wiederkehrende Phänomene können mit einem einzelnen Ticket für das gesamte Buch gemeldet werden; in der Regel ist jedoch die Meldung auf eine einzelne Seite bezogen. Die Fundstelle kann bequem markiert werden und wird automatisch mit den entsprechenden Koordinaten auf dem Digitalisat in die Ticketmeldung übernommen. Auf diese Weise wird den Bearbeitern das Auffinden der entsprechenden Stelle in den XML-Dateien erleichtert.

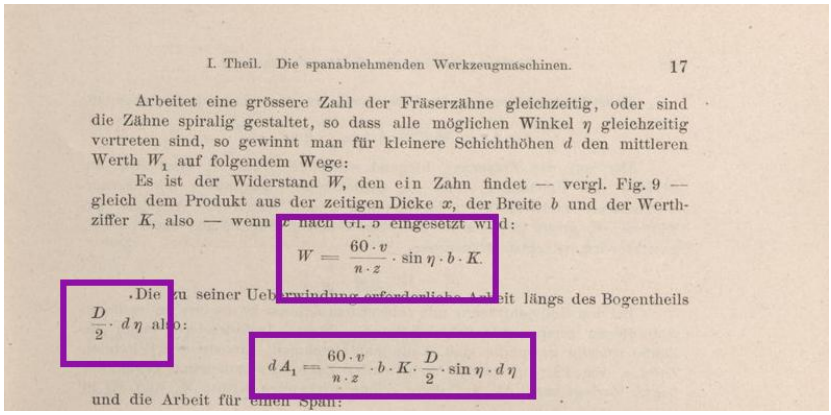
Das Ticketsystem implementiert eine vollwertige und nachnutzbare Historie zur Kategorie, zum Zeitpunkt der Erstellung und Bearbeitung eines Tickets sowie zu dessen Bearbeiter(n). Es ermöglicht so empirisch fundierte Aussagen über Fehler in großen Korpora. Die Meldungen werden an die Bearbeiter des DTA weitergegeben, die diese wiederum im Originaltext überprüfen und beheben, wenn es sich um tatsächliche Fehler handelt.

Entwickelt wird derzeit die Möglichkeit für entsprechend autorisierte Bearbeiter, Korrekturen direkt online umzusetzen. In einer solchen Lösung liegt großes Potential für die breite Anwendung von DTAQ, insbesondere auch für externe Textkorpora. Zudem könnte eine solche Option neben der Fehlerkorrektur auch dazu dienen, (einfache) textkritische Kommentare in TEI-kodierten Texten anzubringen.

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Formeltranskription in DTAQ

Die Option, Korrekturen ad hoc in den Texten umzusetzen, wurde bereits für die Behandlung mathematischer Formeln in DTAQ realisiert. Hierfür existiert ein Werkzeug, welches die Transkription in TeX-Syntax ermöglicht.[13] DTAQ stellt dafür einen Editor mit Vorschauoption zur Verfügung. Formeln aus der Vorlage werden dabei für die Textansicht in Grafiken konvertiert; innerhalb des XML/TEI wird das transkribierte TeX als <formula notation="TeX">[...]</formula> eingebettet.



I. Theil. Die spanabnehmenden Werkzeugmaschinen. 17

Arbeitet eine grössere Zahl der Fräserzähne gleichzeitig, oder sind die Zähne spiralig gestaltet, so dass alle möglichen Winkel η gleichzeitig vertreten sind, so gewinnt man für kleinere Schichthöhen d den mittleren Werth W_1 auf folgendem Wege:

Es ist der Widerstand W, den ein Zahn findet — vergl. Fig. 9 — gleich dem Produkt aus der zeitigen Dicke x, der Breite b und der Werthziffer K, also — wenn x nach Gl. 5 eingesetzt wird:

$$W = \frac{60 \cdot v}{n \cdot z} \cdot \sin \eta \cdot b \cdot K.$$

Die zu seiner Ueberwindung erforderliche Arbeit längs des Bogentheils $\frac{D}{2} \cdot d\eta$ also:

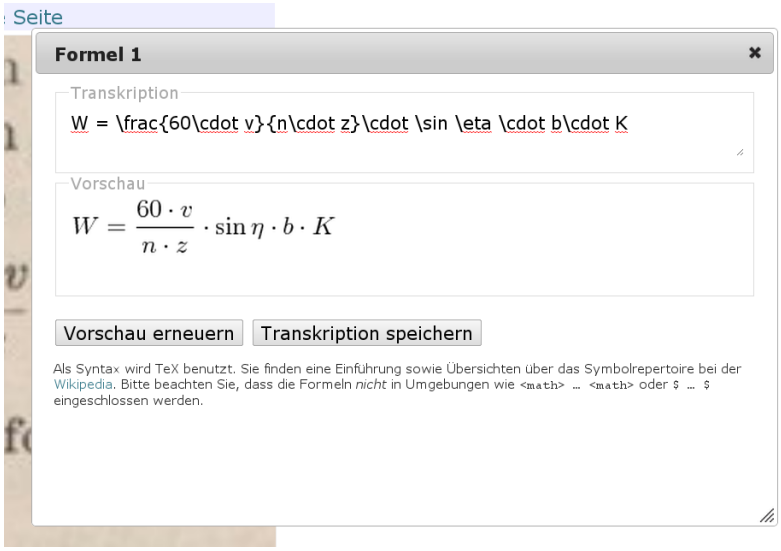
$$dA_1 = \frac{60 \cdot v}{n \cdot z} \cdot b \cdot K \cdot \frac{D}{2} \cdot \sin \eta \cdot d\eta$$

und die Arbeit für einen Span:

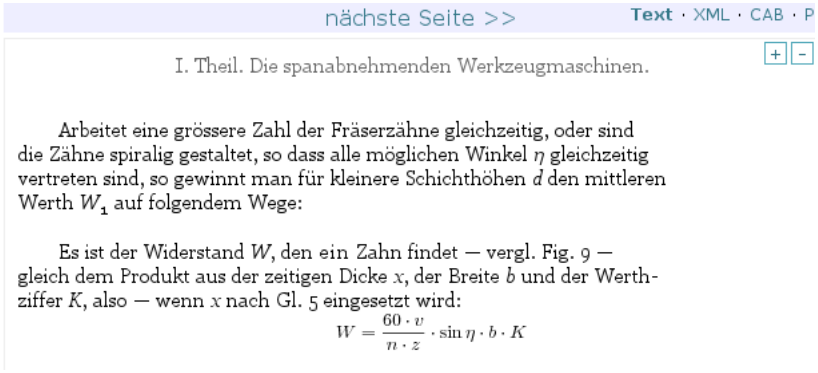
Screenshot: Vorlage mit Formel
Bildausschnitt aus Fischer, Hermann: Die Werkzeugmaschinen. Erster Band: Die Metallbearbeitungs-Maschinen (Textband).
Berlin: Springer, 1900, S. 17. [16]

[13] Vgl. Voß (2010).

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)



Screenshot: Formeleditor



Screenshot: Gerendertes Resultat der Formeleditierung

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Tagging von Druckfehlern

Auch Druckfehler können, sofern sie im Korrekturprozess zuverlässig erkannt werden, direkt in der betreffenden Textstelle annotiert werden. Hierfür wird entsprechend den TEI-Guidelines ein <choice>-Element eingesetzt:

```
<choice>
  <sic>[fehlerhafte Form]</sic>
  <corr>[verbesserte Form]</corr>
</choice>
```

Versionierung von Bearbeitungsständen

Jede Änderung an den XML/TEI-Quellen wird innerhalb eines Repositorys mit Versionierung gespeichert[10] Somit ist jede Version eines Textes stets verfügbar, und alle Änderungen, welche ein Text seit seiner Aufnahme in das DTA erfahren hat, sind protokolliert und mittels diff-Werkzeugen komfortabel nachvollziehbar.

Technische Umsetzung

Die Webanwendung ist in Perl[11] geschrieben (nicht zuletzt aufgrund der herausragenden Unicode-Fähigkeiten) und nutzt, neben vielen Bibliotheken aus dem CPAN[12], eine PostgreSQL-Datenbank als Backend und jQuery[13] für die Implementierung des Benutzer-Interfaces. Sie läuft mit allen gängigen Browsern und ist, neben PCs, auch auf Tablet-Computern und anderen mobilen Geräten voll nutzbar.

Usability – Motivation zur Arbeit mit DTAQ

Eine wichtige Voraussetzung für eine breite Nutzerschaft von DTAQ ist eine einfache und intuitiv einleuchtende Bedienbarkeit der Web-Oberfläche. Das akribische Korrekturlesen zum Auffinden von Tran-

skriptionsfehlern auf Zeichenebene wird durch die direkte Gegenüberstellung von Vorlage und Transkription erleichtert. Selbst Annotations- und Darstellungsfehler sind in der Text-Bild-Darstellung meist auf den ersten Blick erkennbar. Sämtliche Fehler lassen sich sodann mit DTAQ unkompliziert markieren und kategorisieren.
Eine Benutzer- und Rechteverwaltung ist implementiert. DTAQ wurde nach einer dreimonatigen, projektinternen Testphase im Mai 2011 für Mitarbeiterinnen und Mitarbeiter der BBAW freigegeben. Seitdem wurden mehr als 25.000 Tickets angelegt. Seit Ende 2011 ist DTAQ für alle Nutzer des DTA freigeschaltet.

Fazit

Die bisherigen Erfahrungen des DFG-Projekts *Deutsches Textarchiv* zeigen, dass die Qualitätssicherung bei der Volltexterfassung im XML/TEI-P5-Format sowohl formativ als auch summativ erfolgen muss. Das *Double-Keying*-Verfahren ist, insbesondere bei heterogenen Frakturtexten, schlechteren Vorlagen oder komplexeren Textstrukturen hinsichtlich der zeichengenauen Erfassung dem automatisierten OCR-Verfahren deutlich überlegen. Dennoch ist auch das *Double-Keying*-Verfahren in verschiedenen Bereichen fehleranfällig, sodass Maßnahmen zur Qualitätssicherung im Prozess der Volltextdigitalisierung notwendig sind. Um entsprechende Methoden entwickeln und größtmöglich an den Problembereich anpassen zu können, ist die Kenntnis verschiedener Fehlerquellen und Fehlerkategorien essentiell. Erste exemplarische Korrekturgänge an den bislang im DTA enthaltenen Texten ermöglichten eine differenzierte Fehlerklassifikation, die sowohl die Zeichen- als auch die Markupgenauigkeit mit einbezieht. Denn die Genauigkeit bei der Volltexterfassung hängt nicht allein – wie bislang häufig angenommen – von der Messung der Zeichengenauigkeit ab. Vielmehr sind hier weitere Differenzierungen hinsichtlich der Zeichenkodierung und der XML-Annotation vorzusehen.

Für eine Evaluierung der Erfassungs- und Annotationsgenauigkeit auf der Grundlage der hier vorgestellten Fehlerklassifikation führte das DTA eine entsprechende Studie durch, im Rahmen derer mehr als 7000 Seiten der annotierten Volltexte anhand der jeweiligen Vorlage Korrektur gele-

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

sen wurden.[14] Für diesen Arbeitsschritt bot die hier vorgestellte DTA-Korrekturumgebung DTAQ eine nutzerfreundliche verteilte Arbeitsumgebung, die auch in anderen Projektkontexten einsetzbar wäre.

Literatur

Geyken, Alexander

2007 The DWDS Corpus. A reference corpus for the German language of the 20th century. In: Christiane Fellbaum (Hg.): Collocations and Idioms. Linguistic, Lexicographic, and Computational Aspects. London, Continuum Press.

Geyken, Alexander et al.

2011 Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In: Silke Schomburg/Claus Leggewie/Henning Lobin/Cornelius Puschmann (Hgg.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. 20./21. September 2010. Beiträge der Tagung. 2., ergänzte Fassung, Köln, hbz, 157–161 (<http://www.hbz-nrw.de/dokumentencener/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf#page=159>, 24.06.2011).

Holley, Rose

2009 How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. In: D-Lib Magazine 14,3/4 (doi:10.1045/march2009-holley).

Jurish, Bryan

2008 Finding canonical forms for historical German text. In: Angelika Storrer/Alexander Geyken/Alexander Siebert/Kay-Michael Würzner (Hg.): Text Resources and Lexical Knowledge: selected papers from the 9th Conference on Natural Language Processing (KONVENS 2008), Berlin, Mouton de Gruyter, 27–37.

Jurish, Bryan

2010 More than words: using token context to improve canonicalization of historical German. In: Journal for Language Technology and Computational Linguistics, 25, 1, 23–40.

[14] Ergebnisse dieser ersten Studien wurden in einem Vortrag auf der internationalen Konferenz „Philology in the Digital Age: 2011 Annual Conference and Members' Meeting of the TEI Consortium“, Würzburg, 10.–16. Oktober vorgestellt: Geyken, Alexander / Haaf, Susanne: „Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text”. Eine umfassende Auswertung für das Journal of the Text Encoding Initiative ist in Vorbereitung.

Forum Computerphilologie, Geyken/Haaf/Jurish/Schulz/Thomas/Wiegand über *TEI und Textkorpora* <http://.computerphilologie.tu-darmstadt.de/jg09/geykenetal.pdf> (05.08.2012)

Jurish, Bryan
2012 Finite-state Canonicalization Techniques for Historical German. PhD thesis, Universität Potsdam (urn:nbn:de:kobv:517-opus-55789).

Reichmann, Oskar/Klaus-Peter Wegera (Hg.)
1993 Frühneuhochdeutsche Grammatik. Tübingen, Niemeyer.

Unsworth, John
2011 Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI. In: Journal of the Text Encoding Initiative 1 (<http://jtei.revues.org/215>, 24.06.2012).

Tanner, Simon et al.
2009 Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. In: D-Lib Magazine 15,7/8 (doi:10.1045/july2009-munoz).

Voß, Herbert
2010 Math mode, <http://www.tex.ac.uk/tex-archive/info/math/voss/mathmode/Mathmode.pdf> (24.06 2012).

Internetadressen

[1] DTA, Homepage: <http://www.deutschestextarchiv.de> (24.06.2012).
[2] DTA, Basisformat: <http://www.deutschestextarchiv.de/doku/basisformat> (24.06.2012).
[3] TEI, Homepage: <http://www.tei-c.org/> (24.06.2012).
[4] DTA, Richtlinien zur Texterfassung: <http://www.deutschestextarchiv.de/doku/richtlinien> (24.06.2012).
[5] TEI, P5 Guidelines: <http://www.tei-c.org/Guidelines/P5/> (24.06.2012).
[6] DFG, Practical Guidelines on Digitization (Status: April 2009) <http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_en.pdf>
[7] DTA, Parametrisierbare Leseansicht: <http://www.deutschestextarchiv.de/leseansicht> (24.06.2012).
[8] DTA, Qualitätssicherung (DTAQ) <http://www.deutschestextarchiv.de/doku/dtaq> (24.06.2012).
[9] DTAQ, Korrekturfibel: <www.deutschestextarchiv.de/doku/korrekturfibel> (24.06.2012).
[10] Git – Fast Version Control System: <http://git-scm.com/> (24.06.2012).
[11] The Perl Programming Language: <http://www.perl.org/> (24.06.2012).
[12] Comprehensive Perl Archive Network: <http://www.cpan.org/> (24.06.2012).
[13] jQuery – The Write Less, Do More, JavaScript Library: <http://jquery.com/> (24.06.2012).
[14] <http://www.deutschestextarchiv.de/sailer/selbstmord/1785/viewer/image/text/40/183/> (24.06.2012).
[15] <http://www.deutschestextarchiv.de/moritz/nachahmung/1788/viewer/image/text/40/32/> (24.06.2012).
[16] <http://www.deutschestextarchiv.de/fischer/werkzeugmaschinen01/1900/viewer/image/31/ (24.06.2012).