

Edierst Du noch oder trainierst Du schon? Forschungsdaten als Grundlage von Trainingsdaten für die automatische Texterkennung

Boenig, Matthias

boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften,
Deutschland
ORCID: 0000-0003-4615-4753

Baierer, Konstantin

konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz,
Deutschland
ORCID: 0000-0003-2397-242X

Hinrichsen, Lena

hinrichsen@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0002-9286-2390

Würzner, Kay-Michael

kay-michael.wuerzner@slub-dresden.de
Sächsische Landesbibliothek — Staats- und
Universitätsbibliothek Dresden (SLUB), Deutschland
ORCID: 0000-0002-9039-4124

Reul, Christian

christian.reul@uni-wuerzburg.de
Zentrum für Philologie und Digitalität (ZPD) der
Universität Würzburg, Deutschland
ORCID: 0000-0002-1776-1469

Einführung

Wichtigste Grundlage der textorientierten Forschung in den Digital Humanities ist eine ausreichende Verfügbarkeit von hochwertigem maschinenlesbarem Text. Diese Anforderung kann bei grundständig digitalen Texten häufig einfacher erfüllt werden als bei historischen Texten, wo zunächst die Transformation vom gedruckten oder geschriebenen Wort auf Papier in eine geeignete digitale Repräsentation zu realisieren ist.

Mit der Anwendung von Verfahren des maschinellen Lernens in der automatischen Texterkennung ist in den letzten zehn Jahren ein enormer Fortschritt vollzogen worden. Dies betrifft vor allem die Zeichenerkennung und deren Genauigkeit. Hierbei kommen Methoden zum Einsatz, die dem Paradigma *Lernen aus Beispielen* folgen. Die dazu nötigen Trainingsdaten werden als Ground Truth (GT) bezeichnet.

„Der Ursprung des Wortes Ground Truth ist das deutsche Wort Grundwahrheit. Im OCR-Zusammenhang bedeutet das, dass alles auf der gedruckten Seite in gleicher Art und Weise nach definierten Regeln unter anderem in digitaler Form wiedergegeben wird.“¹

Aber GT dient nicht nur dem Training der Zeichenerkennung (sowohl dem Training eines neuen Modells „from scratch“, als auch dem „Finetuning“ eines bestehenden Modells auf einen spezifischen Anwendungsfall hin), sondern wird auch zur Datenvalidierung, -evaluation und -referenzierung eingesetzt. Neben der Zeichenerkennung können aber weitere Teilprozesse der automatischen Texterkennung vom Einsatz maschinellen Lernens profitieren. Dies gilt insbesondere für die Erkennung und Auszeichnung der Seitenstruktur bzw. des Seitenlayouts. Diese unterschiedlichen Anwendungen setzen differenzierte GT-Typen voraus. Allgemein kann zwischen Struktur-GT und Text-GT unterschieden werden.

Die Erstellung von GT erfolgt zu einem Großteil manuell, was einen hohen zeitlichen und finanziellen Aufwand erfordert. Um brauchbaren GT zu erstellen, sind abgestimmte Konventionen und Richtlinien notwendig. Aus diesem Grund entwickelt, pflegt, vermittelt und diskutiert das Projekt OCR-D² neben technischen Lösungen für die Massenvolltexterschließung historischer Drucke vom 16. bis 19. Jahrhundert eigene GT-Richtlinien³. Diese Richtlinien werden in einer offenen, zur kollaborativen Datenkultur verpflichtenden Umgebung erstellt und sollen sicherstellen, dass nachnutzbare Forschungsdaten entstehen sowie der Aufwand in der GT-Erstellung minimiert werden kann.⁴

Forschungsdaten im Kontext des Deutschen Textarchives

Im Rahmen des vorgeschlagenen Workshops soll eine solche offene Datenkultur am Beispiel von Forschungsdaten des Deutschen Textarchivs (DTA)⁵ gemeinsam gelebt und so mittelbar ein wertvoller Beitrag zur Qualität historischer Textkorpora geleistet werden. Die Analyse des DTA vor dem Hintergrund der GT-Erstellung soll den Teilnehmenden zeigen, welche Möglichkeiten (und Grenzen) diese Daten bieten.

Betrachtung des DTA-Datenbestandes

Das DTA wurde im Rahmen eines sprachwissenschaftlich orientierten DFG-Projektes erstellt. Der Kernbestand

besteht aus 1500 Druckpublikationen mit einem Gesamtumfang von 540.000 Seiten. Die Text- und Textsortenauswahl, die zeitliche Spanne des Publikationszeitraumes vom frühen 17. bis frühen 20. Jahrhundert, die Verwendung von Erstausgaben und die vorlagengetreue Transkription kennzeichnen diesen Bestand als Grundlage eines Referenzkorpus der frühneuhochdeutschen Sprache. Die Bereitstellung der digitalen Texte erfolgt sowohl in einem XML-basierten Format als auch als unannotierter Rohtext.

Für die Einschätzung der Nutzbarkeit des DTA als Quelle für GT sind nicht nur die Ergebnisdaten relevant. Ein genauerer Blick auf die einzelnen Etappen des ursprünglichen Datenerfassungsworkflows im DTA zeigt bisher ungenutzte Potenziale der einzelnen Datenstände als Trainingsmaterialien für Text- und Strukturerkennung. Die folgende Abbildung illustriert die beiden grundsätzlichen Wege der Volltexterstellung, die im DTA zur Anwendung kamen: Automatische Texterkennung mit anschließender Nachkorrektur („OCR way“) und manuelle Transkription im Vier-Augen-Prinzip („Double Keying way“).

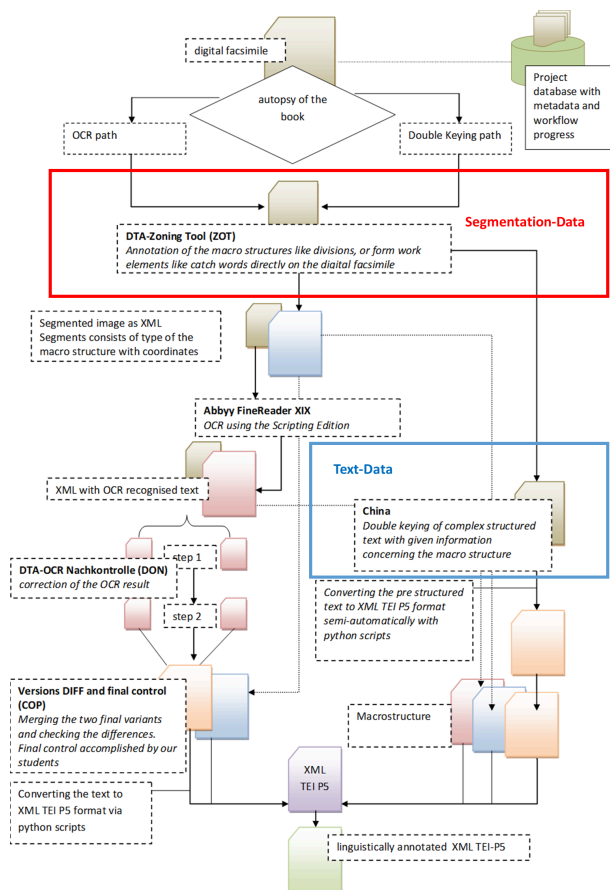


Abbildung 1: Schematische Darstellung des DTA-Datenerfassungsworkflows

Letzterer kam für den Großteil des Bestands zur Anwendung. Das Double-Keying-Verfahren wurde von Nicht-Muttersprachlern vorgenommen und ist sehr genau. Die Zeichengenauigkeit kann mit 99,99 % angesetzt werden (Haaf, 2013; Geyken, 2012). Mit OCR wurden hauptsäch-

lich Titel des 19. und Mitte des 18. Jahrhunderts erfasst.⁶ Für dieses Schrifttum existieren hoch performante und zuverlässige OCR-Modelle.⁷

Beiden Wegen gemein ist ein manueller Segmentierungsschritt. In diesem wurden Textzonen und Abbildungen lokalisiert und klassifiziert. Diese Segmentierung diente zwar „nur“ der nachträglichen Auszeichnung der Volltexte im XML (und nicht etwa der Unterstützung der automatischen Texterfassung). Sie bilden aber dennoch eine der größten bekannten Sammlungen an Strukturdaten für historische deutschsprachige Drucke. Aus der Untersuchung des Datenerfassungsworkflows können somit Segmentierungsdaten und Textdaten identifiziert werden, die für die Verwendung als GT in Frage kämen. Größtes Manko der Datensammlung ist jedoch die fehlende Verknüpfung zwischen Text und Bild, die die Einsatzszenarien als Trainingsdaten massiv einschränkt. An dieser Stelle setzt der vorgeschlagene Workshop an.

Ziel

Die Teilnehmenden des Workshops werden mit Verfahren und Methoden der Erstellung, Erschließung und Speicherung von GT für die automatische Texterkennung vertraut gemacht. Der Workshop ist in zwei Teile geteilt: einen theoretischen und einen praktischen. Ziel des theoretischen Teils ist, dass die Teilnehmenden in die Lage versetzt werden, anhand einer Liste von Kriterien sowie einer Validierung der Daten, Forschungsdaten für die Erstellung von GT einzuschätzen. Mit den OCR-D-GT-Richtlinien bekommen die Teilnehmenden eine in der Praxis erprobte Anleitung für die Erstellung von GT zur Verfügung gestellt. Inhalt und Aufbau, aber auch die Möglichkeiten der praktischen Anwendung dieser Richtlinien im jeweiligen Projekt bilden in diesem ersten Workshopteil den Schwerpunkt. Im praktischen Teil sollen nun die Teilnehmenden in verschiedenen Szenarien GT-Daten erstellen. Dabei werden Forschungsdaten des DTA und vorhandener GT geprüft und eingeschätzt. Dazu werden die im theoretischen Teil vorgestellten Metriken und Validierungsmethoden angewendet. Mit Transformations- und Konvertierungsprogrammen kann in der Folge nun der GT automatisiert erstellt werden. Ebenfalls können spezielle Softwareprogramme für die manuelle Erstellung von GT verwendet werden. Um sich sowohl mit dem Funktionsumfang als auch mit der Leistungsfähigkeit der Tools vertraut zu machen, ist es notwendig, diese im theoretischen Teil kennenzulernen. Der unmittelbare Umgang und die Handhabung des Tools für die GT-Erstellung stehen nicht im Mittelpunkt, sondern die Entscheidung, welches Tool für das jeweilige Vorhaben am geeignetsten scheint.

Zum Abschluss steht die Speicherung des GT in einem Repositoryum. So können die Daten entsprechend der FAIR-Prinzipien zugänglich gemacht werden. Erklärungen zum Aufbau des Repositoryums sowie die Erschließung mit Metadaten, die Nutzung des OCR-D-GT-Repo-Template⁸ schließen diesen Teil und den Workshop ab.

Inhalte

Den Teilnehmenden des Workshops sollen verschiedene Methoden und Verfahren der GT-Erstellung vorgestellt werden.

Theoretischer Teil

1. Prüfung und Bewertung von Forschungsdaten
2. Vorstellung der OCR-D-GT-Richtlinien
3. Vorstellung von Verfahren zur Alignierung von existierenden Transkriptionen und generierter Segmentierung
4. Vorstellung von Softwaretools

Praktischer Teil

1. Erstellung von GT aus Forschungsdaten
 1. Bewertung der Forschungsdaten
 2. Transformation, Konvertierung der Forschungsdaten
 3. Erstellung und Validierung des GT
2. Erstellung des GT durch Transkription
 1. Vorstellung und Anwendung von Transkription-GT-Tools
3. Speicherung und Veröffentlichung des GT
 1. Erstellung eines GT-Repositorys auf GitHub

Benötigte technische Ausstattung:

die jeweiligen Teilnehmenden verfügen über:

- einen GitHub-Account
- Laptop mit installierten Werkzeugen (Liste wird vorab per E-Mail geschickt)
- Ggf. eigene Daten

Der Raum verfügt über:

- Beamer, Whiteboard (wenn möglich)
- Internet via W-Lan

Umfang:

- vier Stunden (90 Minuten Theorie 30 Minuten Pause 120 Minuten Praxis)

Forschungsfeld der Beitragenden

Matthias Boenig ist Informationswissenschaftler sowie Kunsthistoriker. Er hat Bibliotheks- und Informationswissenschaftler und Kunstgeschichte an der Humboldt-Universität zu Berlin studiert. Seit seinem Studium beschäftigt er sich mit der digitalen Transformation von Textdaten in digitale, strukturierte und XML-basierte Forschungsdaten. Dazu war er in verschiedenen Projektkontexten von

1997 an, zu Beginn am Computer- und Medienservice der Humboldt-Universität, dem Institut für Bibliotheks- und Informationswissenschaft und heute an der Berlin-Brandenburgischen Akademie der Wissenschaften tätig. Zurzeit ist Matthias Boenig wissenschaftlicher Mitarbeiter im Projekt OCR-D. Im Rahmen dieses Projekts hat er die OCR-D-GT-Richtlinien entwickelt und betreut diese. Sein derzeitiges praktisches und forschungsorientiertes Interesse besteht in der Erstellung, der Bereitstellung und Standardisierung von GT für die OCR. Matthias Boenig war Mitarbeiter am „Deutschen Textarchiv“.

Kay-Michael Würzner ist Sprachwissenschaftler und hat Computerlinguistik und Germanistik studiert. Nach dem Studium arbeitete er als wissenschaftlicher Mitarbeiter an der Universität Potsdam und der Berlin-Brandenburgischen Akademie der Wissenschaften im Bereich korpuslinguistischer Forschungsdateninfrastrukturen. Seit April 2019 ist Kay-Michael Würzner an der SLUB tätig und bearbeitet Themen des maschinellen Lernens und der automatischen Sprachverarbeitung. Er koordiniert außerdem die Angebote der SLUB rund um einen offenen Forschungskreislauf.

Konstantin Baierer arbeitet seit 2018 als wissenschaftlicher Mitarbeiter für die Staatsbibliothek zu Berlin am Projekt OCR-D, insbesondere an der technischen Interoperabilität der entwickelten Lösungen, der OCR-D/core Softwarebibliothek und dem Release Management. Besonders wichtig sind ihm transparente, inklusive und robuste Methoden für verteilte Softwareentwicklung und gute Schnittstellen zwischen Kulturerbeeinrichtungen und Digital Humanities.

Lena Hinrichsen ist wissenschaftliche Mitarbeiterin an der Herzog August Bibliothek Wolfenbüttel und dort seit 2021 als Koordinatorin im Projekt OCR-D tätig. Ihr Studium der Buchwissenschaft absolvierte sie an der Johannes Gutenberg-Universität Mainz.

Dr. Christian Reul leitet die Digitalisierungseinheit des Zentrums für Philologie und Digitalität (ZPD) der Universität Würzburg. Seine Forschungsschwerpunkte sind die OCR/HTR auf historischem Material sowie die Neu- und Weiterentwicklung der entsprechenden Software.

Fußnoten

1. vgl. OCR-D-GT-Richtlinie < <https://ocr-d.de/de/gt-guidelines/trans/trLevels.html> >
2. DFG-Projekt OCR-D : Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR), Koordinierungsprojekt < <https://ocr-d.de/de/about> >
3. OCR-D-GT-Richtlinien <<https://ocr-d.de/de/gt-guidelines/trans/>>
4. siehe dazu Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke (Engl 2020)
5. Deutsches Textarchiv <<https://deustextarchiv.de/>>

6. siehe dazu Deutsches Textarchiv – Der Digitalisierungsworkflow im DTA < <https://deutschestextarchiv.de/doku/workflow> >
7. Siehe dazu GT4Hist-GT-Datensatz mit Korrekturen: <https://code.bib.uni-mannheim.de/ocr-d/GT4HistOCR>, Training Fraktur from GT4HistOCR: <https://github.com/tesseract-ocr/tesstrain/wiki/GT4HistOCR>, Modelle: GT4HistOCR: <https://code.bib.uni-mannheim.de/ocr-d/GT4HistOCR/src/branch/master/models>, frak2021: <https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/>
8. OCR-D/gt-repo-template: A template for creating a ground truth repo with the various functions and features: such as metadata creation, data analysis and presentation. (github.com) < <https://github.com/OCR-D/gt-repo-template> >

Bibliographie

Engl, Elisabeth; Boenig, Matthias; Baierer, Konstantin; Neudecker, Clemens; Hartmann, Volker. 2020: „Volltexte für die Frühe Neuzeit : Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke“. Zeitschrift für Historische Forschung, 47(2), 223-250. doi:10.3790/zhf.47.2.223.

Haaf, Susanne; Wiegand, Frank; Geyken, Alexander: „Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text“. Journal of the Text Encoding Initiative (jTEI) 4, 2013. doi:10.4000/jtei.739.

Geyken, Alexander; Haaf, Susanne; Jurish, Bryan; Schulz, Matthias; Thomas, Christian; Wiegand, Frank: „TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv“. Jahrbuch für Computerphilologie – online, 2012, <http://computerphilologie.digital-humanities.de/jg09/geykenetal.html>.