Professionalisierung der Ausbildung von Geisteswissenschaftlern in der Digitalisierung von Texten

Dahnke, Michael
fedja_anatevka@web.de
Universität Würzburg, Deutschland

Motivation

»Angesichts der steigenden Sichtbarkeit der Digital Humanities, auch und gerade bei universitären Schwerpunktsetzungen, ist die Frage, wie sie am sinnvollsten gelehrt werden sollen, von steigender Bedeutung« […] »Um diese Bemühungen längerfristig in der Community zu verankern, wurde auf der ersten Jahreskonferenz der Digital Humanities der deutschsprachigen Länder im März 2014« […] »eine Arbeitsgruppe der DHd gegründet. Die Proponenten der Arbeitsgruppe schlugen vor,« […] »die bisher losen Diskussionen stärker auf ein »Referenzcurriculum« zu fokussieren.« (https://dig-hum.de/ag-referenzcurriculum-digital-humanities)

Diesem Anliegen fühlt sich der als Dozent für die Vermittlung von Digitalisierungskompetenz an der Universitätsbibliothek Würzburg arbeitende Autor verpflichtet. Er engagiert sich in der *AG Referenzcurriculum Digital Humanities*, ist Mitglied des Würzburger Arbeitskreis Digitale Editionen und unterstützt das Editionsprojekt *Narragonien digital*. Er plädiert mit seinem Beitrag dafür, die DH-Ausbildung bezüglich der Bilddigitalisierung und des OCR (Optical Character Recognition) stärker zu kanonisieren. Mit seiner eigenen Lehrveranstaltung *Bilddigitalisierung und OCR für Geisteswissenschaftler* , deren Schwerpunkt auf der Erstellung der Digitalisate und dem OCR liegt, bietet er eine Referenz, die er hiermit zur Diskussion in der Community stellt.

Das Ziel der Lehrveranstaltung ist die Vermittlung von Kenntnissen des gesamten Digitalisierungsprozesses für MA-, BA- und LA-Studentinnen und Studenten aller geisteswissenschaftlichen Fachrichtungen. Diese nachfolgend als Zielgruppe Bezeichneten sollen in die Lage versetzt werden, selbständig strukturiert ausgezeichnete, digitale Volltexte zu erzeugen und die Ergebnisse der Erstellung derselben beurteilen zu können. Diese Kenntnisse sind wichtig, weil Projekte häufig auch dadurch gefährdet sind, dass Vertretern der Zielgruppe die mangelnde Qualität der ihnen vorliegenden Digitalisate zu spät bewusst wird. Darum müssen sie von Beginn an Digitalisate auf ihre Brauchbarkeit für die automatische Texterkennung beurteilen können. Strukturiert ausgezeichnete, digitale Volltexte sind unabdingbar beispielsweise für Topic Modeling oder Sentiment Analysis auf größeren Textcorpora und als Zwischenstufe für die Erstellung digitaler Editionen.

Ablauf

Kurz gefaßt sind die sechs Arbeitschritte dafür

1. die Sensibilisierung für juristische Aspekte der Bilddigitalisierung,
2. die Suche nach vorhandenen oder die Erstellung von eigenen Digitalisaten,
3. deren Vorverarbeitung,
4. das OCR,
5. die Anreicherung der Digitalisate und des generierten Rohtextes – im Idealfall einer diplomatische Transkription – mit Metadaten, sowie
6. die inhaltliche Auszeichnung des generierten Rohtextes.

2.1. Rechtliche Grundlagen der Bilddigitalisierung

Um den Teilnehmern den typischen Arbeitsablauf der Textdigitalisierung möglichst stringent und ohne thematische Abschweifungen vorzuführen, werden sie zuerst intensiv mit den juristischen Grundlagen der Bilddigitalisierung vertraut gemacht. Dazu gehören

1. die Vorstellung des UrhG beziehungsweise speziell der § 60d und 60g UrhG in der novellierten Fassung des UrhWissG 2017,
2. die Unterscheidung zwischen Immaterial- und Materialgüterrecht und was daraus für die Digitalisierung zweidimensionaler Objekte folgt,
3. die Persönlichkeitsrechte des Urhebers und weiterer Betroffener sowie
4. der Umgang mit Werken, die unter der Creative Commons Lizenz stehen und die Möglichkeit, diese selbst zu benutzen.

2.2. Suche nach vorhandenen Digitalisaten beziehungsweise deren Erstellung

Am Anfang der Transformation vom gedruckten zum digitalen Corpus steht die Suche nach möglicherweise bereits vorhandenen Digitalisaten. Diese Suche setzt neben nicht DH-spezifischen Kenntnissen der Erschließung das Wissen um Metadaten zu digitalen Bildformaten voraus. Die Vertreter der Zielgruppe müssen in dieser Situation wissen, dass beispielsweise die Chancen eines erfolgreichen OCR mit einem JPEG mit 72 dpi deutlich

geringer sind als mit einem unkomprimierten TIFF, True Color und 300 dpi. Sollte er schließlich feststellen, dass ihm Digitalisate in der gewünschten Form nicht zugänglich sind, bedarf er der Kenntnisse zu digitalen Bildformaten genauso, um im nächsten Schritt erfolgreich den Scan selbst durchzuführen oder nach seinen Vorgaben durchführen zu lassen.

Ausgehend von den skizzierten Anforderungen werden den Vertretern der Zielgruppe in der Veranstaltung die Grundlagen der Bilddigitalisierung nahe gebracht. Vertieft wird hier auf das menschliche Sehen und die Farbreproduktion, die Entstehung digitaler Bilder (Rastergraphik, optische und interpolierte Auflösung, Farbtiefe), Farbräume, Color-Management-Systeme, verschiedene Graphikspeicherformate, Speichermedien und verschiedene Scannertypen eingegangen. Auch für die Berücksichtigung konservatorischer Aspekte werden die Teilnehmer sensibilisiert.

2.3. Aufbereitung der Digitalisate für das OCR

Die Forschung im Bereich OCR, insbesondere auf Inkunabeln und Wiegendrucken, sowie die eigene Praxis des Autors belegen die Bedeutung einer vorherigen Aufbereitung der Digitalisate für das OCR, der darum entsprechend Platz in der Veranstaltung eingeräumt wird (Springmann 2015: 9). Als Tätigkeiten sind hier in der Reihenfolge ihrer Ausführung die Bereinigung und anschließende Binarisierung der Digitalisate, deren Segmentierung in einzelne Textabschnitte und schließlich einzelne Textzeilen sowie die Transkription (›ground truth‹) einer Anzahl der Textzeilen zu nennen. Die gesamte Aufbereitung der Digitalisate für das OCR führen die Vertreter der Zielgruppe in der Veranstaltung selbst an ausgewähltem Trainingsmaterial durch. Nach der Teilnahme an der Veranstaltung sollen sie auch diesen Arbeitsschritt selbständig erledigen und Arbeitsergebnisse anderer in diesem Bereich beurteilen können.

2.4. OCR

Entsprechend der zunehmenden Spezialisierung des Digitalisierungszentrums der UB Würzburg ist es erstens wünschenswert, in der Lehrveranstaltung besonders auf das Training eigener Modelle beispielsweise mit *OCRopus* einzugehen. Zweitens soll ein Arbeitsablauf für die Digitalisierung eigener Texte vorgestellt werden, der von den Vertretern der Zielgruppe selbständig mit möglichst geringem technischen Aufwand realisierbar ist. Diese Form der Digitalisierung von Texten soll als handhabbares Mittel zum Zweck wahrgenommen werden.

Schließlich wird in diesem Zusammenhang auf die Frage nach dem Zeitpunkt der Normalisierung des mit dem OCR erstellten Textes eingegangen. Soll bereits mit dem OCR ein normalisierter Text erstellt werden und wenn ja, nach welchen Regeln? Ist also beispielsweise von der verwendeten OCR-Software das Schaft-s bereits automatisch als Rund-s zu lernen und anschließend zu transkribieren? Oder soll das Ergebnis des OCR graphisch so dicht wie möglich am Original bleiben und normalisierende Eingriffe erst hinterher erfolgen?

2.5. Auszeichnung/Anreicherung

Nach dem OCR wird erst die Notwendigkeit der Auszeichnung sowohl der Digitalisate mit Metadaten als auch des Rohtextes erläutert, die die Teilnehmer dann auch selbst vornehmen sollen. Für den extrahierten Rohtext gilt das in zweifacher Hinsicht: Erstens sind ihm Metadaten hinzuzufügen, welche die spätere, eindeutige Identifikation des Werkes und dessen Auffindbarkeit ermöglichen. Zweitens muss der Text strukturiert mit inhaltsbezogenen Elementen angereichert werden.

Konkret sind bei der digitalen Repräsentation eines Romans beispielsweise die Figuren, Orte, Zeitpunkte und gegebenenfalls weitere signifikante Entitäten im Text für das spätere, automatisierte Retrieval zu kodieren. Andere Anforderungen stellen digitalierte Transkriptionen gesprochener Sprache und wiederum andere die Erstellung einer Urkundenedition (Vogeler 2015). Für die visuelle Präsentation, beispielsweise auf einem Webportal, sind textstrukturierende Merkmale wie die Einteilung nach Kapiteln, Abschnitten, Fußnoten etc. zu kennzeichnen. Dem unterschiedlichen Kenntnisstand der Teilnehmer geschuldet muss hier vor der Vorstellung der TEI Guidelines zuvor zweifelsohne XML dargestellt werden. Wie ausführlich daneben Dublin Core und bibliotheksspezifische Formate (MARC21) thematisiert werden, ist noch nicht entschieden. Wieviel Zeit für weiterführende Themen wie Named Entity Recognition , PID und Normdaten wie GND bleibt, muß ebenfalls die Praxis weisen.

Forschungsbezug und Weiterentwicklung

Die eine Stärke der Würzburger Lehrveranstaltung ist die Praxisorientierung, der nach der zweitägigen Einführung noch stärker mit einer drei Tage dauernden Übung Rechnung getragen wird. Bei dieser werden die Vertreter der Zielgruppe mit vorbereiteten Scans selbständig die genannten Arbeitsschritte von der Bereinigung und anschließenden Binarisierung über die Segmentierung und Transkription, dem OCR bis zum anschließenden Auszeichnen beziehungsweise der Anreicherung des digitalen Corpus mit den nötigen Metadaten vornehmen.

Unverzichtbar für die gesamte Ausbildung im DH-Bereich ist neben dem Praxisbezug die Orientierung am neuesten Stand der Forschung in allen Teilbereichen. Dem wird bei der skizzierten Lehrveranstaltung erstens durch die Forschung einzelner Mitglieder

des Digitalisierungszentrums als die Veranstaltung verantwortende Abteilung Rechnung getragen (1. Reul/ Wick/Springmann/Puppe: 2017. 2. Springmann: 2016. 459–462). Zweitens ist die enge Zusammenarbeit des Digitalisierungszentrums mit dem Lehrstuhl für Informatik VI der Universität Würzburg zu nennen.

Denkbare Erweiterungen für eine zukünftig breiter angelegte Lehrveranstaltung der beschriebenen Art sind die Vorstellung a) des OCR von Handschriften, beispielsweise in der Kooperation mit einer Transkribus anwendenden Institution, idealerweise der *Digitalisierung und elektronische Archivierung – DEA* der Universität Innsbruck, und b) des Einsatzes virtueller Forschungsumgebungen zur Herstellung digitaler Ressourcen.

Aus Sicht des Autors ist nach der erfolgreichen Durchführung die kritische Diskussion mit den Autors ähnlicher oder gleicher Veranstaltungen von anderen Institutionen unverzichtbar. Ausgehend von http://cceh.uni-koeln.de/digitale-geisteswissenschaften-studiengange-2011/ befragt er aktuell die Mitarbeiter einschlägiger Institutionen nach deren Angeboten im Bereich der Digitalisierung von Texten. Er hofft mit seinem Beitrag wie dem AG Treffen an der DHd2018 auf einen fruchtbaren Meinungsaustausch.

Bibliographie

Corbach, Almuth: *Bestandsschonendes Digitalisieren von schriftlichem Kulturgut*. In: Digital und analog. Die beiden Archivwelten. 46. Rheinischer Archivtag. Ratingen 21.-22. Juni 2012.

Jannidis, Fotis / Hubertus Kohle / Malte Rehbein [Hrsg.]: *Digital Humanities. Eine Einführung*. Springer-Verlag GmbH Deutschland, 2017.

Kneißl, Michael: *Scannen wie die Profis : Text- und Bildvorlagen perfekt digitalisieren*. München: DTV. (2)2002.

Loewenheim, Ulrich / Adolf Dietz / Gerhard Schricker: *Urheberrecht*. Kommentar. München: Beck. (4)2010.

Reul, Christian / Christoph Wick / Uwe Springmann / Frank Puppe: *Transfer Learning for OCRopus Model Training on Early Printed Books*. In: *Zeitschrift für Bibliothekskultur*. Bd. 5, Nr. 1 (2017).

Springmann, Uwe: *A high accuracy OCR method to convert early printings into digital text. A Tutorial*. *Center for Information and Language Processing (CIS)*. LMU. München. 2015. S. 9.

Springmann, Uwe: *OCR für alte Drucke*. *Informatik-Spektrum*. 39(6):459–462. 2016.

Vogeler, Georg: *Die Text Encoding Initiative (TEI) als Werkzeug des Urkundeneditors – Erfahrungen und Desiderate*. In: Fees, Irmgard Prof. Dr.; Hotz, Benedikt; Schönfeld, Benjamin (Hrsg.): *Papsturkundenforschung zwischen internationaler Vernetzung und Digitalisierung. Neue Zugangsweisen zur europäischen Schriftgeschichte*. Göttingen. 2015.

Weitzmann, John H. / Paul Klimpel: Rechtliche Rahmenbedingungen für Digitalisierungsprojekte von Gedächtnisinstitutionen. Berlin: Zuse Institute Berlin. digiS – Servicestelle Digitalisierung Berlin. (3)2016.