Best of Both Worlds
Zur Kombination algorithmischer und manueller Verfahren bei der Erschließung großer Handschriftenkorpora

Nantke, Julia
julia.nantke@uni-hamburg.de
Universität Hamburg, Germany

Bläß, Sandra
sandra.blaess@uni-hamburg.de
Universität Hamburg, Germany

Flueh, Marie
marie.flueh@uni-hamburg.de
Universität Hamburg, Germany

Maus, David
david.maus@sub.uni-hamburg.de
Staats- und Universitätsbibliothek Hamburg, Germany

Einleitung

Für historische Zeiträume vor der Etablierung elektronischer Kommunikationsformen bilden Briefe das zentrale Medium des Austauschs über räumliche Distanzen hinweg. Aus heutiger Sicht sind Briefe deshalb wichtige historische Quellen, die persönliche Beziehungen ebenso dokumentieren wie gesellschaftlich relevante Orte, Ereignisse, Themen etc. Dies gilt umso mehr für ein solch umfangreiches Korrespondenznetz wie jenes des Ehepaars Ida und Richard Dehmel. Die Dehmels bildeten um 1900 das Zentrum eines europaweiten Netzwerks von Künstler:innen und Kulturschaffenden. An der in ca. 35.000 Briefen überlieferten Korrespondenz waren viele der zentralen Akteur:innen des damaligen Kunst- und Kulturbetriebs beteiligt.

Begreift man Erinnerung mit Jan Assmann als kreativen Gestaltungsprozess (vgl. Assmann 1999: 16), so lässt sich zunächst grundsätzlich feststellen, dass digitale Verfahren der Erschließung und Präsentation eine andere „Formung“ (Assmann 1999: 32) kultureller Vergangenheit ermöglichen als analoge, maßgeblich gedruckte Formate. In Bezug auf das Briefnetzwerk der Dehmels ermöglicht die Digitalisierung durch neue Verfahren der computergestützten Erschließung und Präsentation eine Konzeptualisierung und Darstellung des Briefnetzwerks als kulturhistorischer Zusammenhang. Diese Forschungsperspektive rekurriert auf die Fähigkeit insbesondere schriftlicher Alltagsdokumente, Dialoge, Gedanken und Diskurse als „Zeitinseln“ (Assmann 1988: 12) zu transportieren und für die Nachwelt zu konservieren. Im Zuge einer kuratierenden Erschließung lassen sich diese Zeitinseln entsprechend für die wissenschaftliche und kulturhistorische Rezeption in der Gegenwart repräsentieren. Als „immutable mobiles“ (Latour 1990: 26) zeugen die Briefe zudem nicht nur von zeithistorischen Ereignissen und gesellschaftlichen Entwicklungen des europäischen kulturellen Lebens um 1900, sondern in ihnen materialisieren sich auch die mit dem Medium Brief in dieser Zeit verknüpften kulturellen Praktiken und Kommunikationsformen. Nicht zuletzt zeigt außerdem der Fall des zu Lebzeiten weltberühmten und nach seinem Tod 1920 rasant dekanonisierten Dichters Richard Dehmel eindrücklich, wie stark literarische Moden und personell-institutionelle Konstellationen von kulturbetrieblichen, gesellschaftlichen und politischen Dynamiken abhängen und sich im Laufe der Zeit verändern.

Ziel des Projekts *Dehmel digital* ist es, die im Dehmel-Archiv der Staats- und Universitätsbibliothek Hamburg (SUB) archivierten Briefe in eine digitale Repräsentation zu überführen, welche die in den Briefen gespeicherten persönlichen, kulturellen und gesellschaftlichen Dynamiken erfahrbar und erforschbar macht. Kulturhistorische Zusammenhänge werden anhand ihres konkreten textuellen Niederschlags in den brieflichen Quellen erschlossen und repräsentiert (vgl. dazu auch Baßler 2005: 176f.; Baillot 2011). Indem die Briefe im Projekt als digitale Quellen rekontextualisiert und publiziert werden, erhalten sie gleichzeitig einen neuen „Ort der kanonisierten Erinnerung“, von dem aus sie vergegenwärtigt und erinnert werden und so zum kulturellen Gedächtnis beitragen können (vgl. Assmann 1999: 31).

Im Sinne des kulturellen Gedächtnisses ist das Briefkorpus vor allem als Ganzes interessant und relevant: Denn um nicht nur Richard Dehmels subjektive Wahrnehmung zum Kulturdiskurs der Jahrhundertwende zu erfassen, ist es unerlässlich, sich nicht auf dessen eigene Zeitzeugnisse zu beschränken, sondern auf das gesamte Gruppengedächtnis seines Netzwerks zurückzugreifen. Ein konzeptueller Wandel digitaler Editionsformate zugunsten der Dokumentation personeller und institutioneller Zusammenhänge deutet sich bereits an, steckt aber editionspraktisch noch in den Anfängen (Nutt-Kofoth 2020; Nantke 2019). Die Plattform *Briefe und Texte aus dem intellektuellen Berlin um 1800*, die *digitale Quellenedition Der Sturm* sowie die Edition *Jean Paul – Sämtliche Briefe digital* sind teilweise noch in der Entwicklung befindliche Editionsprojekte, deren Materialauswahl und -präsentation anstelle von Einzelautor:innen und deren Werken auf kommunikative und institutionelle Netzwerke ausgerichtet sind. Sie bilden in ihrer Anlage Vorbilder für das Projekt *Dehmel digital*.

Zentral für die Etablierung solcher Editionsformate ist nicht zuletzt, dass den neuen digitalen Möglichkeiten der umfänglicheren *Repräsentation* personeller, institutioneller und medialer Netzwerke, die in den genannten Beispielen erprobt werden, auch entsprechende Verfahren der computergestützten *Erschließung* zur Seite gestellt werden. Dabei ist es entscheidend, Möglichkeiten und Grenzen einer computationellen Erschließung großer Handschriftenkorpora zu reflektieren, die dem erhöhten Materialumfang, den es zu erschließen gilt, gerecht werden. Es gilt zu diskutieren, in welchem Verhältnis Erschließungsumfang und textkritische Prüfung stehen sollen und wie sich die aktuellen digitalen Möglichkeiten nutzen lassen, um computergestützt auch große Datenmengen zu bewältigen, Zusammenhänge darzustellen und auf diese Weise neue Orte der Erinnerung zu etablieren. Welche computationellen Verfahren lassen sich an welchen Schnittstellen miteinander kombinieren und mit welchen Limitationen ist dabei zu rechnen? Wie kann also in Anbetracht begrenzter personeller und zeitlicher Ressourcen eine gute Mitte zwischen quantitativ-statistischen Verfahren und den qualitativ-philologischen Anforderungen einer digitalen Repräsentation gefunden werden? Für das Projekt *Dehmel digital* gilt, dass die Erfassung und adäquate Darstellung dieser großen Menge an Quellen nur im Rahmen einer Kombination algorithmischer quantitativer Verfahren und manueller Praktiken umsetzbar ist. Wir verfolgen deshalb

den Ansatz, die Qualitäten und arbeitspraktischen Vorteile beider Welten gewinnbringend zu verbinden.

Der Workflow

Der Prozess der Überführung der handschriftlichen Originale in maschinenlesbare Repräsentationen untergliedert sich in eine Reihe aufeinander aufbauender Transformationen, die miteinander zusammenhängende Abstraktionsschichten vom Ursprungsmaterial produzieren (vgl. Abb. 1).

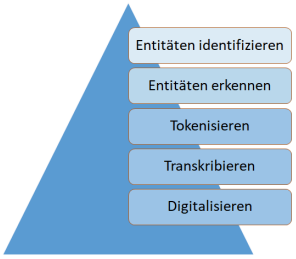

Abb. 1: Abstraktionsschichten des Workflows

Dabei besteht der Anspruch, ein Verfahren zu etablieren, welches zum einen die Erschließung eines möglichst großen Teils der Briefe ermöglicht, aber zum anderen einer editionsphilologisch validen Repräsentation der Dokumente verpflichtet bleibt. Entsprechend dieser doppelten Zielsetzung greifen in dem im Rahmen des Projekts entwickelten Workflow manuelle Arbeitsschritte und algorithmisch getriebene Prozesse ineinander. Dabei kombinieren wir mehrere bereits innerhalb der Digital Humanities etablierte Verfahren, die wir für den Einsatz in einer Edition modifizieren (vgl. Abb. 2).

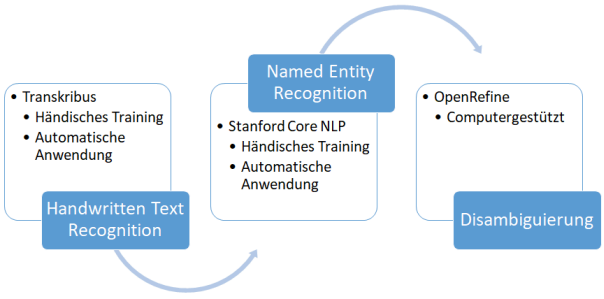

Abb. 2: Die einzelnen Arbeitsschritte

Nachdem die Originalbriefe in der SUB in hochauflösende Bilddigitalisate übersetzt und mit grundlegenden Metadaten angereichert wurden, transkribieren wir zunächst einige Briefe einer:eines Schreibenden manuell in *Transkribus* (vgl. https://readcoop.eu/de/transkribus/), bis genug Daten als Basis für das Training eines HTR-Modells erzeugt wurden. Im Rahmen dieses manuellen Schritts werden zudem in strukturierter Form weitere Metadaten zu den Briefen erfasst. Nach erfolgreichem Training eines HTR-Modells in *Transkribus* erfolgt die weitere Transkription im iterativen Wechsel zwischen automatisierter Transkription und manueller Nachkorrektur, wobei die korrigierten Briefe wiederum als Trainingsdaten in den HTR-Workflow eingespeist werden. Die Transkriptionen werden als PAGE XML aus *Transkribus* exportiert.

Eine XML-Verarbeitungskette wandelt diese Transkriptionen in ein TEI-basiertes Format um und wendet von uns trainierte Modelle des *Stanford Named Entity Recognizers* (vgl. Manning et al. 2014) an, um Personen, Orte, Institutionen und Werke zu taggen. Da sich die klassischen Entitäten in Privatbriefen aus der Zeit um 1900 anders darstellen als in Gebrauchstexten, anhand derer die meisten Classifier trainiert wurden, nutzen wir Teile unserer Transkriptionen sowie die Daten weiterer digitaler Briefeditionen für das Training eigener Classifier. Ergänzend zum überwachten maschinellen Lernen werden Listen implementiert, um sicherzustellen, dass bestimmte, regelmäßig wiederkehrende Entitäten zuverlässig gefunden werden. Die sukzessive Erweiterung des Trainingsmaterials führt zunächst zu besseren Erkennungsraten innerhalb unseres Korpus, ist perspektivisch aber auch als generische Möglichkeit der automatisierten Briefannotation gedacht, die im Sinne einer digitalen, kollaborativen, wissenschaftlichen Korrespondenzanalyse auch für andere Projekte zugänglich gemacht werden soll. Erkannte Entitäten werden *inline* unter Beibehaltung der Bezüge zu Layout und Struktur des Dokuments eingebracht. Hierfür wird eine vereinfachte Implementierung der *Separated Markup API for XML* (vgl. Verwer 2020) verwendet, die klassifizierte Tokens der Named Entity Recognition und textstrukturelles Markup integriert. Die so erkannten Entitäten werden als Basis für die teilautomatisierte Generierung von Personen-, Orts-, Institutionen- und Werkregistern genutzt. Erst dieser Schritt einer basalen inhaltlichen Erschließung ermöglicht letztlich einen grundlegenden Überblick über sowie gezielte Einblicke in das umfangreiche Korpus anhand zentraler Entitäten und damit eine Nutzung als editorisch aufbereitete historische Quelle.

Um die wiederum anhand maschineller Lernverfahren teilautomatisiert ermittelten Entitäten in stabile Registereinträge zu überführen, wird ein maschinengestützter Abgleich (data reconciliation) mit Normdatensystemen und lokalen Wissensbasen durchgeführt. Die mit *OpenRefine* (vgl. https://openrefine.org/) computergestützt disambiguierten Entitäten-Typen bilden wiederum die Basis für manuell zu erstellende Makrokommentare zu den zentralen Akteur:innen, Institutionen und Werken des Korpus sowie für Netzwerkvisualisierungen, die wiederum mit den Briefen verlinkt sind und einen vom grafisch visualisierten Netzwerk ausgehenden Einstieg in die textnahe Lektüre der Briefe ermöglichen. Darüber hinaus werden die disambiguierten Entitäten mit Normdaten zu Personen, Orten, Werken und Organisationen verknüpft und die Briefe des Dehmel-Korpus via API in den Webservice correspSearch (https://correspsearch.net/de/start.html) eingebunden, mit dessen Hilfe Verzeichnisse verschiedener Briefeditionen nach Absender, Empfänger, Schreibort und -datum durchsucht werden können. Auf diese Weise wird das um die Dehmels bestehende Korrespondenznetz in bereits etablierte Netzwerkzusammenhänge integriert.

Ergebnis dieser – im Vortrag anhand von konkreten Beispielen genauer darzustellenden – Reihe von Datentransformationen sind also transkribierte, annotierte und mit Metadaten angereicherte XML-Dateien sowie stabile, disambiguierte und mit Normdaten verknüpfte Entitätenregister, die gemeinsam den Input für das Webportal von *Dehmel digital* bilden.

Die Sicherung und Pflege der Daten übernimmt die SUB Hamburg, sodass die nachhaltige Verfügbarkeit und Nutzbarkeit unserer Ergebnisse gewährleistet sind.

Repräsentation: Nutzungsszenarien und Gestaltungsfragen im Hinblick auf ein ‚digitales Gedächtnis'

Die eingangs beschriebene Relevanz des Korpus als Teil eines digital repräsentierten Gedächtnisses europäischer Kulturgeschichte ist mit der Frage einer adäquaten Repräsentation verknüpft. Hierbei sind Überlegungen im Hinblick auf die Kontextualisierung der Inhalte relevant, die bereits durch die Anlage des Erschließungsworkflows vorstrukturiert werden: Im Projekt *Dehmel digital* steht anstelle der Äußerungen von Einzelpersonen die Korrespondenz als Netzwerkzusammenhang im Fokus; die im Dehmel-Archiv konservierten einzelnen Zeitinseln sollen verstärkt in einen Kontext zueinander und zu den Inhalten anderer digitaler Editionen gebracht werden können. Die materielle Erschließung, semantische Annotation und inhaltliche Kommentierung erfolgen deshalb aus der Perspektive einer möglichst breiten, dezentralen Dokumentation des Korrespondenznetzes als historisches Zeugnis eines kollektiven Gedächtnisses.

Dementsprechend ist es das Ziel des im Rahmen von *Dehmel digital* entwickelten Webportals, möglichst vielfältige Nutzungsszenarien von der skalierbaren Lektüre (vgl. Weitin 2017) bis hin zu maschinell gestützten Auswertungen anhand eigener Forschungsfragen zu ermöglichen. Hierbei sind unterschiedliche Nutzer:innengruppen von kulturinteressierten Museumsbesucher:innen des Dehmelhauses (vgl. https://www.dehmelhaus.de/aktuell.html) bis hin zur (digital arbeitenden) geisteswissenschaftlichen Community mitgedacht. Latour zufolge erreicht Wissen viele Menschen an verschiedenen Orten am wirkungsvollsten, wenn es in mobiler, beständiger, präsentierbarer, lesbarer und kombinierbarer Form vorliegt (vgl. Latour 1990: 23–26). Eine digitale Präsentation auf einem Webportal erweist sich insbesondere in Bezug auf die Darstellbarkeit von Netzwerken und die Mobilität der Präsentation im Allgemeinen als besonders effektiv, da sie dauerhaft von überall aus kostenlos mit eigenen Geräten abgerufen werden kann, nicht an ephemere Materialien gebunden ist und verschiedene Aufbereitungen je nach Zielgruppe und Nutzungsinteresse miteinander kombiniert werden können (vgl. dazu grundsätzlich bezogen auf digitale Repräsentationen auch Sahle 2016: 30): Auf dem Portal selbst unterstützen facettierte Suchen, Netzwerk- sowie Kartenvisualisierungen und Makrokommentare die strukturierte Rezeption sowie eigenständige Recherchen im Korpus. Sie bilden moderierte Einstiege in das umfangreiche Material, die sich bei Bedarf an individuelle Interessen flexibel anpassen lassen (vgl. Spoerhase 2015: 640–643). Über das Portal hinaus stehen die produzierten Daten der Nachnutzung in anderen Forschungsszenarien offen. In diesem Sinne werden nicht nur die Faksimiles und die erzeugten Transkriptionen in verschiedenen Formaten (TEI, Plaintext, PDF), sondern ebenfalls die für den beschriebenen Workflow entwickelten HTR- und NER-Modelle sowie unsere Routinen zum Download zur Verfügung gestellt.

Die Digitalisierung des kulturellen Gedächtnisses ermöglicht in der Kombination manueller und algorithmischer Arbeitsprozesse zum einen überhaupt erst die quantitativ-qualitative Erschließung des Dehmelschen Korrespondenznetzes. Zum anderen ist die digitale Repräsentation ebenfalls die Bedingung für die vielfältigen und skalierbaren Rezeptionsszenarien desselben als kulturelles Artefakt – und zwar nicht als Zeugnis einer isolierten Vergangenheit, sondern einer, die durch Diskussion und individuelle Aneignung der Quellen mit der Gegenwart verbunden werden kann (vgl. Assmann 1988: 13).

Der Beitrag stellt den hier beschriebenen Workflow anhand von konkreten Beispielen dar und gibt anhand der ersten öffentlichen Beta-Version Einblicke in die geplante und bereits prototypisch implementierte Umsetzung auf dem Portal. Dabei wird es ebenfalls darum gehen, anhand der Beispiele den Umgang mit dem Spannungsfeld zwischen quantitativer Erschließung und klassischer philologischer Arbeit zu diskutieren.

Bibliographie

Assmann, Jan (1988): "Kollektives Gedächtnis und kulturelle Identität". In: Ders.; Hölscher, Tonio (Hrsg.): *Kultur und Gedächtnis*. Frankfurt am Main, S. 9–19.

Assmann, Jan (1999): "Kollektives und kulturelles Gedächtnis. Zur Phänomenologie und Funktion von Gegen-Erinnerung". In: Borsdorf, Ulrich; Grütter, Heinrich Theodor (Hrsg.): *Orte der Erinnerung. Denkmal, Gedenkstätte, Museum*. Frankfurt am Main/ New York, S.13–32.

Baillot, Anne (2011): "Einleitung". In: Dies. (Hrsg.): *Netzwerke des Wissens. Das intellektuelle Berlin um 1800*. Berlin 2011, S. 11–23.

Baßler, Moritz (2005): *Die kulturpoetische Funktion und das Archiv. Eine literaturwissenschaftliche Text-Kontext-Theorie*. Tübingen.

Briefe und Texte aus dem intellektuellen Berlin um 1800 (o.D.): Hrsg. v. Anne Baillot, Humboldt Universität Berlin. URL: http://www.berliner-intellektuelle.eu/.

Der Sturm. Digitale Quellenedition zur Geschichte der internationalen Avantgarde (2018): Hrsgg. von Marjam Trautmann und Torsten Schrade, Mainz, Akademie der Wissenschaften und der Literatur. URL: https://sturm-edition.de.

Jean Paul. Sämtliche Briefe digital (2018): Hrsgg. im Auftrag der Berlin-Brandenburgischen Akademie der Wissenschaften von Markus Bernauer, Norbert Miller und Frederike Neuber. URL: https://www.jeanpaul-edition.de/start.html.

Latour, Bruno (1990): "Drawing things together". In: Michael E. Lynch und Steve Woolgar (Hrsg.): *Representations in Scientific Practice*. Cambridge, S. 19–68.

Manning, Christopher / Surdeanu, Mihau / Bauer, John / Finkel, Jenny / Bethard, Steven J. / McClosky, David (2014): "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, S.55–60.

Nantke, Julia (2019): "Konzepte digitaler (Re-)Präsentationen von Literatur zwischen Pluralisierung und Standardisierung". In: Martin Huber, Sybille Krämer und Claus Pias (Hrsg.): *Forschungsinfrastrukturen in den digitalen Geisteswissenschaften. Wie verändern digitale Infrastrukturen die Praxis der Geisteswissenschaften?* Frankfurt a. M., S. 58–76. urn:nbn:de:hebis:30:3-526104

Nutt-Kofoth, Rüdiger (2020): "Der Brief als Forschungsfeld - Editionswissenschaft". In: Marie Isabel Matthews-Schlinzig, Jörg Schuster, Gesa Steinbrink und Jochen Strobel (Hrsg.): *Handbuch Brief. Von der Frühen Neuzeit bis zur Gegenwart*. Berlin/Boston, S. 81–96.

Sahle, Patrick (2016): "What is a Scholarly Digital Edition?" In: Matthew James Driscoll und Elena Pierazzo (hrsg.): *Digital Scholarly Editing: Theory and Practice*. Cambridge, S. 19–39. DOI: http://dx.doi.org/10.11647/OBP.0095.02.

Spoerhase, Carlos (2015): "Gegen Denken? Über die Praxis der Philologie". In: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 89, S. 637–646.

Verwer, Nico (2020): "Plain text processing in structured documents". In: *Proceedings of Declarative Amsterdam 2020*. DOI: 10.1075/da.2020.verwer.plain-text-processing.

Weitin, Thomas (2017): "Scalable Reading". In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47, S. 1–6.

o.V.: *corresp Search*. URL: https://correspsearch.net/de/start.html [13.Juli 2021].

o.V.: Dehmelhaus Stiftung Hamburg. URL: https://www.dehmelhaus.de/aktuell.html [29. November 2021].

o.V.: *OpenRefine*. URL: https://openrefine.org/ [13. Juli 2021].

o.V.: *Transkribus. KI-gestützte Handschriftenerkennung*. URL: https://readcoop.eu/de/transkribus/ [13. Juli 2021].