BmE

Separatum aus:

THEMENHEFT 12

Elisabeth Lienert / Joachim Hamm
Albrecht Hausmann / Gabriel Viehhauser (Hrsg.)

Digitale Mediävistik

Perspektiven der Digital Humanities für die Altgermanistik

Publiziert im November 2022.

http://www.erzaehlforschung.de – Kontakt: herausgeber@erzaehlforschung.de
ISSN 2568-9967

Zitiervorschlag für diesen Beitrag:
Viehhauser, Gabriel: Digitalisierung von Handschriften und frühen Drucken, OCR (Bericht über Kurzvorstellungen und Diskussion Sektion 1), in: Lienert, Elisabeth/Hamm, Joachim/Hausmann, Albrecht/Viehhauser, Gabriel (Hrsg.): Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik, Oldenburg 2022 (BmE Themenheft 12), S. 17–## (online).

Gabriel Viehhauser

Digitalisierung von Handschriften und frühen Drucken, OCR (Bericht über Kurzvorstellungen und Diskussion Sektion 1)

Die automatische Erkennung von Handschriften und Drucken gehört wohl zu jenen Bereichen auf dem Gebiet der digitalen Methodik, in denen die Fortschritte, die in den letzten Jahren erzielt worden sind, am offenkundigsten (und auch für Laien ersichtlich) ins Auge fallen: Ähnlich wie bei der ebenso in jüngerer Zeit merkbar verbesserten Übersetzungssoftware werden diese Fortschritte weit über den akademischen Bereich hinaus wahrgenommen und sind letztlich dem *data-* bzw. *deep-learning-turn* geschuldet. Erst der Einsatz von elaborierten maschinellen Lernverfahren, denen große Mengen an Trainingsmaterial zugrunde liegen, hat hier zu den entscheidenden Verbesserungen im Vergleich zu früheren Versuchen geführt.

Für die Mediävistik ist die Handschriften- und Frühdruckerkennung neben ihrer praktischen Bedeutung von nicht zu unterschätzender grundlegender Signifikanz und hätte durchaus das Potential, einen Paradigmenwechsel in der Vorstellung von mittelalterlicher Schriftlichkeit an sich hervorzurufen: So erscheint es nun absolut in Greifweite, Handschriften in großer Zahl im Volltext zu digitalisieren und auch bisher unbeachtete Texte (oder Textvarianten) für literaturwissenschaftliche, linguistische und kulturwissenschaftliche Auswertungen zugänglich zu machen. Damit ergibt sich die Chance, auch solche Texte ins wissenschaftliche Blickfeld zu bringen, deren manuelle Erschließung bislang als zu aufwendig oder nicht

lohnend erschien. Bekanntermaßen beruhen solche Aufwandseinschätzungen nicht selten auf ästhetischen Werturteilen und berühren Fragen der Kanonbildung. Auch wenn es zu diskutieren bleibt, ob Kanonisierungsprozesse durch einen *Distant Reading*-Zugang im Sinne Franco Morettis (Moretti 2016) überwunden werden können (bzw. überhaupt sollten), so dürfte alleine die bloße Möglichkeit der Verfügbarkeit einer großen Masse von Texten zu Verwerfungen im Feld kanonischer Selbstverständlichkeiten führen. Der in der digitalen Editorik bereits ersichtliche Trend zur Überlieferungsnähe (und die sich daran anschließenden Debatten, inwieweit diese auch literaturwissenschaftlich relevant ist), dürfte sich jedenfalls, bedingt durch die Verbesserungen auf dem Gebiet der Handschriftenerkennung, weiter fortsetzen.

Zugleich zeichnet sich angesichts der Möglichkeiten dieser Verfahren aber auch ein Paradigmenwechsel oder, weniger radikal gedacht, eine Erweiterung im Spektrum philologischer Zugangsweisen ab: Automatische Handschriftenerkennung wird selbst bei besttrainierten Modellen zu einer gewissen Fehlerrate in der Erkennung führen. Ohne nachträgliche manuelle Korrektur (*post-processing*) wird also kein philologisch einwandfreier Text entstehen. Es stellt sich nun die Frage, ob sich die Altgermanistik darauf einlassen kann und will, schnell und kostengünstig automatisch erstellte, aber in Details womöglich fehlerhafte Textversionen zu akzeptieren, natürlich nicht als unbesehen übernommene Grundlage für eine philologisch exakte Edition, aber doch ergänzend zur überblicksmäßigen Auswertung und zur sinngemäßen Erschließung von Textkonvoluten, bei denen es vielleicht nicht in jedem Aspekt auf hundertprozentige Genauigkeit in der Textwiedergabe ankommt; oder aber als Ausgangsbasis für Volltextsuchen, bei denen dann ergebnisbezogen Fehler an Einzelstellen im Transkriptionstext verbessert werden können.

Auch wenn man nicht so weit gehen möchte, so bringen die Fortschritte in der Handschriften- und Frühdruckerkennen jedenfalls unbestreitbar neue Potentiale für die Texterschließung und Zugänglichmachung mit sich

und können die Arbeit etwa an digitalen Editionen und Textrepositorien erheblich erleichtern bzw. kostengünstiger machen. In der entsprechenden Sektion auf unserer Tagung wurden zwei aktuelle Systeme im Bereich der Handschriften- und Frühdruckerkennung vorgestellt, die sich besonders für den Einsatz in der Altgermanistik eignen und bereits in einigen Projekten zur Anwendung kommen, nämlich Transkribus und OCR4all.

Im Vortrag von Günter Mühlberger zu Transkribus standen insbesondere dessen Geschichte, das gegenwärtige Geschäftsmodell und typische Workflows im Vordergrund. Mühlberger berichtete von den Anfängen des OCR in der Frakturerkennung und davon, dass in den 2000er-Jahren noch ›traditionelle‹ Formen des OCR vorherrschten, bis dann der *data turn* das Feld revolutioniert habe. Ab 2013 habe auch das Vorgängerprojekt von Transkribus (tranScriptorium) *machine-learning*-Methoden zum Einsatz gebracht, mittlerweile sei das System in der Digital-Humanities-Szene (und darüber hinaus) stark verbreitet: Laut einer Auswertung von in der Fachliteratur erwähnten Tools gehörte Transkribus 2019 zu den meistverwendeten Programmen in den digitalen Geisteswissenschaften (vgl. https://weltliteratur.net/dh-tools-used-in-research/). 77000 User seien bislang registriert, pro Tag arbeiteten derzeit ca. 100 Nutzer und Nutzerinnen mit dem Werkzeug. Diese Zahl sei stetig am Steigen und wachse über den engeren Kreis der Digital Humanities hinaus (etwa in kommerzielle Bereiche oder in das Gebiet der Ahnenforschung).

Aus dieser Sachlage ergäben sich Anforderungen, die ab 2013/14 zur Erstellung eines neuen, in den Digital Humanities noch wenig verbreiteten Geschäftsmodells geführt haben, nämlich zur Einrichtung einer Genossenschaft. Diese Genossenschaft, die heute 102 (auch außereuropäische) Teilnehmer zählt, arbeite zwar profitorientiert, aber vor allem für den Selbsterhalt; es gebe in diesem Sinne keinen *shareholder value*. In Rechnung gestellt werde insbesondere die Texterkennung, dazu würden Einnahmen durch Kundenprojekte lukriert.

Transkribus stelle große, vortrainierte Modelle zur Verfügung, daneben könnten eigene Modelle trainiert werden, die auch öffentlich freigegeben werden könnten - aber nicht müssten. Mehr als 12.000 Modelle seien bereits von Nutzer und Nutzerinnen erstellt worden, der Schwerpunkt liege allerdings eher im 19. Jahrhundert und nicht in der Mediävistik.

Zum Abschluss seines Vortrags skizzierte Mühlberger Empfehlungen für Workflows bei der Anwendung von Transkribus. So sollten am Anfang der Projektarbeit umfangreiche Vorüberlegungen dazu angestellt werden, welches Ergebnis letztendlich erzielt werden soll: Welcher Grad an Genauigkeit der Transkription wird erwartet, ist eine Strukturauszeichnung mit TEI geplant bzw. wie tief soll diese erfolgen und schließlich, als Grundlage für all diese Entscheidungen, wieviel Aufwand kann und soll für die Erstellung der Umschriften betrieben werden? Sollen spezielle Zeichen und Abkürzungen verwendet werden (neuerdings ermöglicht Transkribus auch das Mittrainieren von Abkürzungszeichen und deren Auflösungen)? Wie viele Handschriften bzw. Drucke umfasst das Projekt und wie viel Text beinhalten diese jeweils? Vor der Arbeit sollte jedenfalls überprüft werden, ob öffentliche Modelle zugänglich bzw. deren Einsatz für die gewünschte Transkriptionsqualität ausreichend sind. Das selbständige Trainieren eines auf die eigenen Anforderungen spezialisierten Modells beginne sich bei Einzelmanuskripten ab 100 Seiten Länge zu lohnen.

In der an den Vortrag anschließenden Diskussion (Leitung: Freimut Löser) standen insbesondere Fragen zu den Modellen im Vordergrund, etwa welche speziellen Modelle (z. B. für Koberger) es bereits gebe (Martin Schubert), ob sich Modelle über unterschiedliche Editionsprojekte hinweg aggregieren ließen (Michael Stolz; aufgrund von überschneidenden Richtlinien ist dies problematisch) und ab wann es sinnvoll werde, ein eigenes Modell zu veröffentlichen (Gabriel Viehhauser; sinnvoll ab ca. 100.000 trainierten Wörtern). Bereits hier wurde deutlich, dass es wünschenswert wäre, den Austausch von mediävistischen Modellen durch verstärkte Koordination innerhalb des Fachs zu befördern. Schließlich richtete sich eine

Frage auf die Einsatzbarkeit von Transkribus für ideographische Sprachen (Meihui Yu; alle horizontal angeordneten Schriftzeichen können mittrainiert werden).

Im Vortrag von Christian Reul zu OCR4all standen vor allem arbeitspraktische Aspekte im Vordergrund. Im Gegensatz zu Transkribus ist OCR4all eine Open-Source-Software, die durch die unbeschränkte Replizierbarkeit nachhaltig und ausbaubar bleibe, dieser Vorteil werde aber durch die nicht-triviale Entwicklung erkauft. OCR4all soll insbesondere eine niederschwellige Plattform zur Verfügung stellen, welche in einer Live-Demo präsentiert wurde. Das Tool bietet neben der Schrifterkennung vor allem eine interaktive Oberfläche für die Nach-Korrektur des maschinell erkannten Textes. Maschinelle Erkennung und manuelles *post-processing* gehen damit Hand in Hand. Auch hier zeige sich, dass beim Einsatz von digitalen Methoden insbesondere durch die Kombination von automatischen und manuellen Verfahren die besten Ergebnisse erzielt werden können und Interfaces, die diesen Zusammenhang berücksichtigen, einen besonderen Nutzen erbringen.

In Hinblick auf die Leitfrage der Tagung („Was braucht das Fach?“) gab Reul zu bedenken, dass dies mitunter dem Fach selbst nicht klar sei. Jedenfalls sei immer ein *trade-off* zwischen Qualitätsanspruch und Aufwand zu berücksichtigen. Nicht zuletzt, um Verluste dieses *trade-offs* zu minimieren, solle, wenn möglich, zumindest auf gemischte Modelle zurückgegriffen werden, also eigenes Training von vorhandenen Modellen ausgehen.

Auch dieser Vortrag führte somit letztlich auf das Desiderat der Zusammenarbeit bei der Modellerstellung, denn je mehr Modelle vorhanden und für die Community verfügbar sind, desto positiver fällt die Kosten-Nutzen-Rechnung beim Einsatz von Schrifterkennungssystemen aus. In der allgemeinen Diskussion der beiden Vorträge wurde daher zunächst nach der Austauschbarkeit der trainierten Modelle zwischen Transkribus und OCR4all gefragt (Viehhauser). Zwar sind die beiden Systeme verschieden und in der Tiefe nicht kompatibel, Mühlberger wies jedoch darauf

hin, dass die zum Training erstellten Daten (also Digitalfaksimiles und deren zeilengenaue manuelle Umschrift) freilich identisch seien und ausgetauscht werden könnten.

Einen weiteren Diskussionspunkt stellte die Frage dar, ob sich die Systeme durch Einbeziehung von linguistischen Daten (wie etwa durch Wörterbuchabgleich oder Informationen über die Sprachstruktur) verbessern ließen (Elisabeth Lienert, Albrecht Hausmann). Automatische Nachkorrekturen schienen aber eher zu Verschlimmbesserung zu führen; auch der Einsatz von *transformer*-Sprachmodellen, die Kontextwahrscheinlichkeiten berücksichtigen, führe derzeit nur zu geringen Verbesserungen.

Damit war in der Diskussion schließlich die Frage nach dem Perfektionsgrad erreicht, der mit Hilfe von automatischen Verfahren angestrebt werden sollte: Inwieweit ist es sinnvoll, die mittlerweile ohnedies hohen Erkennungsraten aufwändig noch weiter zu verbessern, oder sollte nicht eher das Augenmerk auf die Verbesserung von Nachkorrektur-Tools gesetzt werden (Joachim Hamm)? Dass solche *post-correction*-Prozesse sinnvollerweise mit Erkennungsalgorithmen zusammengedacht werden sollten, demonstriere ja insbesondere das OCR4all-Projekt. Kurt Gärtner wies darauf hin, dass sich eine solche Postkorrektur auf alle Fälle lohne, und Andrea Rapp plädierte dafür, mehr Mut zur Publikation von Zwischenergebnissen zu zeigen, auch wenn diese nicht perfekt sind. Freimut Löser erinnerte daran, dass Transkribus (ebenso wie OCR4all) den Vorteil biete, die Texttranskription mit dem Digitalfaksimile zeilengetreu zu verbinden (was die Stellen in ihrer korrekten handschriftlichen Textgestalt leicht auffindbar macht). Ein fehlerfreier, perfekter Text sei ohnedies nie zu erreichen, und es wäre zu erwägen, ob die rasche Verfügbarmachung von großen Textmassen für die Volltextsuche einen perfekten Text überhaupt voraussetze. Unter Umständen könne also eine solche rasche Verfügbarmachung sinnvoller sein. Stephan Müller wies in diesem Zusammenhang darauf hin, dass die DFG Texterkennungsprojekte nach durchaus unterschiedlichen Kriterien für förderungswürdig erachtet. Wichtig sei jedenfalls, Güte-

klassen und Fehlerquotienten anzugeben. Auch Hausmann schloss sich dem Plädoyer für die Fehlertoleranz mit dem Hinweis an, dass für korpuslinguistische Explorationen Rohdaten und Textmassen ausschlaggebender seien als perfekte Editionen. Simone Schultz-Balluff berichtete, dass sie bei ihren eigenen Editionsprojekten (wie von Rapp angeregt) Daten schnell öffentlich zur Verfügung gestellt habe; diese Daten seien auch mit semantischen Auszeichnungen versehen worden.

Die Diskussion wurde schließlich durch Hinweise zu weiterführenden Perspektiven abgerundet: Jakub Simek fragte danach, wie sich Transkriptions-Tools in einen Editions-Workflow einbinden ließen. Mühlberger schlug die Integration über eine API vor und wies darauf hin, dass sich die Qualität der Erkennung nicht nur aufgrund der Datenmengen, sondern auch aufgrund technischer Weiterentwicklung in den nächsten zehn Jahren nochmals verbessern werde (woraus sich die Frage ergibt, ob zum Teil bereits automatisch erfasste Texte nochmals erschlossen werden sollen). Torsten Schaßan stellte schließlich das Projekt einer umfassenden digitalen Erschließung sämtlicher Handschriften, die sich im Besitz öffentlicher Einrichtungen befinden, im Rahmen der Handschriftenzentren in Aussicht und fragte danach, ob dabei eine Volltexterfassung von der Community gewünscht sei bzw. wie eine solche geplant und wo die Ergebnisse gespeichert werden sollten. Eine solche Perspektive macht nochmals deutlich, welches Potential sich aus der Anwendung digitaler Methoden im Bereich der Handschriften- und Frühdruckerkennung für die Altgermanistik ergeben könnte.

Literaturverzeichnis

Sekundärliteratur

Moretti, Franco: Distant Reading, Konstanz 2016.

Online-Ressourcen

OCR4all: http://www.ocr4all.org/.
tranScriptorium: https://cordis.europa.eu/project/id/600707.
Transkribus: https://readcoop.eu/transkribus/.
weltliteratur.net: https://weltliteratur.net/dh-tools-used-in-research/.

Anschrift des Berichterstatters:

Prof. Dr. Gabriel Viehhauser
Universität Stuttgart
Institut für Literaturwissenschaft
Herdweg 51
70174 Stuttgart
E-Mail: viehhauser@ilw.uni-stuttgart.de