Historische Wetterdaten im Spannungsfeld von OCR und UCD

Constantin Lehenmeier[1], Manuel Burghardt[2]

Universitätsbibliothek Regensburg[1]
Computational Humanities, Universität Leipzig[2]

Zusammenfassung

Dieser Beitrag beschreibt informatische Herausforderungen im Kontext eines Digital Humanities-Projekts zur Erschließung und Analyse historischer Wetteraufzeichnungen im Zeitraum 1774 - 1827. Bei der Erschließung der handschriftlichen Aufzeichnungen, die Besonderheiten wie numerische Messwerte in Tabellenstruktur und überlagernde Notizen enthalten, soll langfristig ein entsprechend trainierter OCR-Ansatz (*optical character recognition*) zum Einsatz kommen. Für die Erstellung entsprechender Trainingsdaten sowie für die manuelle Korrektur der automatisch erkannten Daten ergeben sich zunächst softwareergonomische Herausforderungen aus Perspektive der Medieninformatik. Der Fokus dieses Beitrags liegt daher auf der Erstellung von Tools unter Berücksichtigung von Prinzipien des *usability engineering* und des *user-centered design* (UCD) für geisteswissenschaftliche Forschungsvorhaben.

1 Einleitung: Erschließung historischer Wetterdaten

In Zusammenarbeit mit der Universitätsbibliothek Regensburg soll ein unikaler Bestand bisher unveröffentlichter Wetteraufzeichnungen im Zeitraum 1774 - 1827 des Klosters Sankt Emmeram (Regensburg) erschlossen werden.[1] Die meteorologischen Aufzeichnungen kennzeichnen sich durch ein hohes Maß an Homogenität und Kontinuität und eignen sich damit in besonderer Weise für systematische, computergestützte Analyseansätze im Sinne der Digital Humanities. Die digitalen Daten sollen später von unterschiedlichen geisteswissenschaftlichen Disziplinen zur effektiveren und schnelleren Untersuchung bestehender oder neuartiger Fragestellungen verwendet werden. Die rekonstruierte Darstellung von meteorologischen Momentaufnahmen kann insbesondere Geographen und Historikern helfen gesellschaftliche Auswirkungen und Folgen von Klimaänderungen auszumachen (vgl. Allan et al. 2016, S. 166). Zusätzlich können die Wetterdaten durch weitere regionale Archivbestände ergänzt und deren

[1] Weitere Informationen zum Projektkontext unter http://www.bibliothek.uni-regensburg.de/meteorologie/index.html; Hinweis: alle URLs in diesem Beitrag wurden zuletzt überprüft am 3.7.2018

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2018 in
M. Burghardt & C. Müller-Birn (Hrsg.): INF-DH 2018 – Workshopband, 25. Sept. 2018, Berlin
 https://doi.org/10.18420/infdh2018-10

Zusammenwirken zur effektiven Aufarbeitung lokalhistorischer Entwicklungen und Ereignisse herangezogen werden. So können die gemessenen Werte zur Erklärung schwankender Getreidepreis dienen, die in Rechnungsbüchern der gleichen Zeitperiode zu finden sind.

Vor der computergestützten Analyse der Wetterdaten steht zunächst die Digitalisierung und Erschließung derselben. Die zeit- und kostenaufwendige Erstellung maschinenlesbarer Transkriptionen der historischen Dokumente soll durch automatische OCR-Systeme unterstützt werden. Die in den Wettertagebüchern enthaltenen Tabellen, Zahlen und Symbole können durch gängige OCR-Tools wie bspw. dem *ABBYY FineReader*[2] oder *Transcribus*[3] noch nicht zufriedenstellend erkannt werden. Durch den Einsatz neuster Forschungserkenntnisse im maschinellen Lernen soll künftig die Erkennung spezieller Dokumentstrukturen weiterentwickelt werden: Rekurrente neuronale Netze können etwa dabei helfen unterschiedlichen Schreibern, großer Unleserlichkeit und beschädigten Dokumente entgegenzuwirken (vgl. Zhan et al. 2017). *Convolutional Neural Networks* erzielen vielversprechende Ergebnisse bei der Erkennung von Tabellen (vgl. Oliveira et al. 2018).

Somit bestehen einerseits informatische Herausforderungen im Bereich der OCR von numerischen und tabellarischen Wetterdaten, andererseits soll die Erstellung von Analyse- und Visualisierungstools den Umgang mit den gewonnen Daten unterstützten und den Erkenntnisgewinn befördern. Der Fokus dieses Beitrags liegt auf Herausforderungen bei der Erkennung handschriftlicher Daten und dabei in besonderem Maße auf softwareergonomischen Aspekten entsprechender OCR-Tools.

2 OCR und der Faktor Usability

Die Usability, also die Benutzerfreundlichkeit, mag im Kontext automatischer OCR-Tools zunächst verwundern. Jedoch sind bereits bei der Erstellung von Trainingsdaten Nutzerinteraktionen mit den entsprechenden OCR-Systemen notwendig. Bei einer mangelhaften Erkennung können Trainingsdatensätze helfen, die automatische Texterkennung des Tools zu verbessern und müssen daher besonders sorgfältig und genau vom User erstellt werden. Da Fehler trotzdem nie auszuschließen sind, sollten Korrekturen und Ergänzungen genauso effektiv eingearbeitet werden können. Die Benutzerfreundlichkeit ist somit ein übergreifendes Ziel, das alle Aspekte der Anwendung betrifft. Bestehende Tools zur Erkennung und Analyse historischer Dokumente scheinen selten nutzerorientiert entwickelt worden zu sein, was zur Frustration und letztendlich zur Abwendung des Users führt (vgl. Fronhöfer und Mühlbauer 2017, S. 426). Unzureichende Berücksichtigung spezieller Bedürfnisse sowie fachübergreifende Anforderungen eines externen Publikums und ein lückenhafter oder fehlender Design-Prozess in der Softwareentwicklung führen häufig zu einer unzureichende Bedienbarkeit im Kontext von Digital Humanities-Tools (vgl. Burghardt und Wolff 2014, S. 1). Um die Nutzung digitaler Werkzeuge zu begünstigen ist ein nutzerzentriertes Design derselbigen notwendig und somit

[2] ABBYY FineReader: https://www.abbyy.com/de-de/finereader/

[3] Transcribus: https://transkribus.eu/Transkribus/

auch ein praxisnaher Diskurs über die Einführung geeigneter Usability-Methoden in den Digital Humanities. Denn für eine erfolgreiche Adaption, nicht nur bei Fachexperten, ist laut Potts (2015, S. 255) die Priorisierung der Gebrauchstauglichkeit und des Nutzererlebens unverzichtbar.

3 UCD-Prozess für die Erstellung von OCR-Tools

Wie Nutzerbedürfnisse erkannt und die Usability und User Experience im Entwicklungsprozess berücksichtig werden können, beschreibt Levy (2015): (1) Durch die Erstellung von Personas sollen die Motivationen, Bedürfnisse und Ziele potenzieller Nutzer grundlegend erfasst werden. (2) Da es sich dabei zunächst um bloße Annahmen der Entwickler handelt, werden die Personas durch anschließend geführte Interviews mit echten Nutzern überarbeitet und komplementiert. (3) Anschließend soll eine Marktanalyse über die Funktionalitäten existierender Produkte informieren und aktuelle Standards herausgearbeitet werden. Um die Usability bestehender Tools systematisch analysieren sowie Optimierungen in die eigene Entwicklung miteinfließen lassen zu können, bietet sich (4) ein sogenanntes *Heuristic Markup* als Evaluationsmethodik an (vgl. Buley 2013, S. 136). Dabei wird versucht, typische Aufgaben mit dem zu testenden Produkt zu erledigen während man die eigenen Gedanken und Reaktionen sowie Verstöße gegen etablierte Design- und Usability-Standards aufzeichnet.

Dementsprechend wurde im Gespräch mit Archivaren, Editoren, Bibliotheksangestellten und geisteswissenschaftlichen Lehrenden sowie Studierende die Erstellung einer Transkription historischer Dokumente schrittweise erörtert und dabei auf Probleme und verwendete Hilfsmittel eingegangen. Um zusätzlich zu den Nutzeranforderungen einen grundlegenden Überblick zu Basisfunktionalitäten von OCR-Tools zu erlangen, wurden bestehende Tools (s.o.) evaluiert. Hierbei wurde deutlich, dass mentale Modelle der einschlägigen Nutzergruppe und deren domänenspezifische Fachsprache kaum Berücksichtigung im Design finden. Ebenso liefern die Tools keine Orientierungspunkte und Instruktionen, die Nutzer bspw. durch die in den Interviews deutlich gewordenen Arbeitsschritte leiten. In Abbildung 1 ist ein beispielhafter Screenshot der Anwendung dargestellt, der zeigt, wie eine Transkription anhand eines Dokuments erstellt werden kann. Eine verständliche Sprache und eine dem Prozess angepasst Menüführung sollen dem User den Einstieg erleichtern und ihn durch das Programm führen. Die gewonnenen Erkenntnisse sowie die daraus resultierenden verbesserten Designkonzepte wurden in einem frühen Projektstadium zunächst als Sketches umgesetzt. Diese Entwürfe bilden die Grundlage für die Erstellung eines interaktiven Prototypen, der in regelmäßigen Abständen durch Nutzerinterviews und -tests evaluiert werden soll, um Verbesserungsvorschläge frühzeitig einarbeiten zu können.

4 Fazit

Der Beitrag verdeutlicht die Bandbreite an Herausforderungen, die sich aus Perspektive der Informatik in einem Digital Humanities-Projekt ergeben. Neben Spezialanforderungen durch

besondere Strukturelemente und Symbole der historischen Wetterdaten für einen OCR-Ansatz ergibt sich als weitere zentrale Anforderung die Konzipierung einer benutzerfreundlichen Toolkomponente. Auf Vorkenntnisse und Praktiken von Anwendern aus geisteswissenschaftlichen Fachdomänen soll Rücksicht genommen werden, ohne dabei aber ein Tool ausschließlich für Experten zu entwerfen. Die eingesetzten „Guerilla"-Methoden erweisen sich als zeit- und kostengünstig und lieferten im bisherigen Umfang vielversprechende Ergebnisse mit dem Ziel einer nutzerzentrierten Softwareentwicklung.

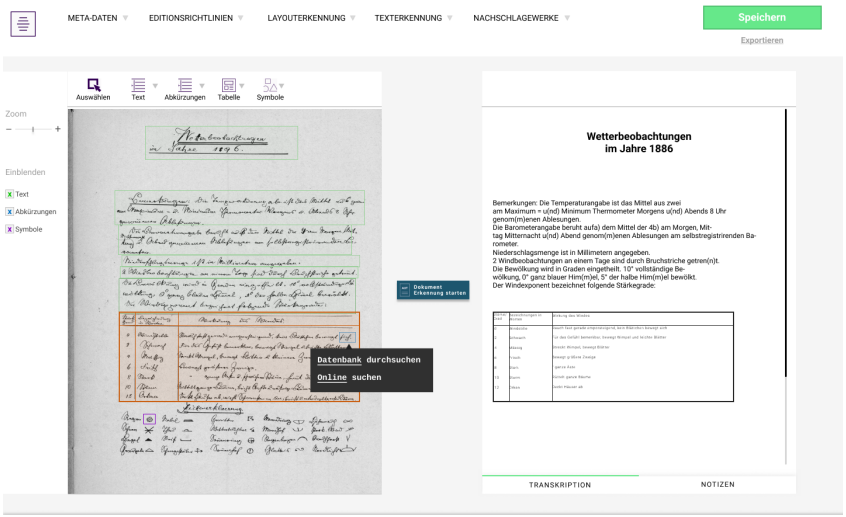

Abbildung 1: Eine übersichtliche und hierarchische Strukturierung der Funktionen, wie im Menü zu sehen ist, soll den Workflow der Nutzer widerspiegeln. Aus dem Dokument (links) werden die erkannten Elemente in die Transkription (rechts) übertragen und Anpassungen eingearbeitet. Die Arbeit soll durch integrierte Online-Suchmaschinen und Datenbanken zu Abkürzungen effizienter gestaltet werden.

Literaturverzeichnis

Allan, R.; Endfield, G.; Damodaran, V.; Adamson, G.; Hannaford, M.; Carroll, F.; Macdonald, N.; Groom, N.; Jones, J.; Williamson, F.; Hendy, E.; Holper, P.; Arroyo-Mora, J. P.; Hughes, L.; Bickers, R.; & Bliuc, A. (2016): Toward integrated historical climate research: the example of Atmospheric Circulation Reconstructions over the Earth. *WIREs Clim Change*, 7 (2), 164–174.

Buley, L. (2013): *The User Experience Team of One. A Research and Design Survival Guide*: Rosenfeld Media.

Burghardt, M.; Wolff, C. (2014): Humanist-Computer Interaction. Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik.

Fronhöfer, A.; Mühlbauer, E. (2017): Archivnutzung ohne Limit. Digitalisierung, Onlinestellung und das Projekt READ für Barrierefreies Forschen. In: *ARCHIVAR,* 70 (4), 422–427.

Levy, J. (2015): *UX strategy. How to devise innovative digital products that people want*. Peking: O'Reilly.

Oliveira, S. A.; Seguin, B.; & Kaplan, F. (2018): dhSegment. A generic deep-learning approach for document segmentation. In: *IEEE Transactions on Cybernetics,* 48 (3), 929–940.

Potts, L. (2015): Archive Experiences: A Vision for User-Centered Design in the Digital Humanities. In Hart-Davidson; W.; Ridolfo; J. (Ed.): *Rhetoric and the Digital Humanities* (255–263). Chicago: University of Chicago Press

Zhan, H.; Wang, Q.; Lu, Y. (2017): Handwritten digit string recognition by combination of residual network and RNN-CTC. In: Derong, L.; Shengli, X.; Yuanqing, L.; Dongbin, Z.; & El-Sayed, E. (Ed.): *Neural Information Processing. 24th International Conference on Neural Information Processing.* Guangzhou, 14.-18.11.2017. Basel: Springer International Publishing, 583–591.