Bericht über den Discourse Lab-Workshop „Zeitungs- und Zeitschriftenkorpora des 19. Jahrhunderts“ am 24.11. in Darmstadt

- David Glück -

I. Zeitungen, Zeitschriften und Novellen im 19. Jahrhundert als Untersuchungsgegenstände

Zunächst werden Motive und Ziele der Digitalisierung von Zeitungen und eventuell Zeitschriften beim Aufbau eines entsprechenden Korpus für das Discourse Lab (Marcus Müller/Jörn Stegmeier) vorgestellt und erläutert. Eine anschließende Präsentation zu überregionalen Tageszeitungen im 19. Jahrhundert (David Glück) stellt sechs relevante Zeitungen heraus und beschreibt deren digitale Verfügbarkeit. Dabei wird einerseits die wechselhafte Pressegeschichte des 19. Jahrhunderts mit ihren (vermeintlichen) historischen Zäsuren deutlich, durch die eine geradlinige Kontinuität von Zeitungen infrage gestellt werden kann; andererseits zeigt sich die mangelhafte digitale Verfügbarkeit der genannten Zeitungen, die unvollständig und „nur“ als Scans für das 19. Jahrhundert verfügbar sind.

Im Anschluss an die Präsentation wird zunächst die Wichtigkeit der Digitalisierung von *überregionalen* Tageszeitungen, die *kontinuierlich* erscheinen, herausgestellt; dies vor allem im Hinblick auf diachrone Analysen und die notwendige Vergleichbarkeit von Zeiträumen bei solchen Analysen. Besonders die *Vossische Zeitung* und die *Frankfurter Zeitung* werden als mögliche Beispiele solcher Zeitungen genannt, andererseits werden auch andere Zeitungen diskutiert, die – wie etwa die *Rheinische Zeitung* – zwar nicht kontinuierlich erschienen sind, aber dennoch einen großen Einfluss ausgeübt haben. Der Betrachtung von kontinuierlich erscheinenden Zeitungen über den gesamten Zeitraum des 19. Jahrhunderts wird die potenzielle Betrachtungweise einer in kleineren, sequenziellen Zeitabschnitten (wie z.B. Dekaden oder Perioden zwischen historischen Zäsuren) operierenden Analyse entgegengehalten, um etwa dem Problem der Diskontinuität von Zeitungen aufgrund historischer Zäsuren (etwa der Märzrevolution 1848) gerecht zu werden. Dieser Argumentation kann jedoch entgegnet werden, dass viele vermeintliche historische Zäsuren bei der Entwicklung von Zeitungen durch die Pressegeschichtsschreibung festgelegt wurden, sich aber durch Analysen in anderen Disziplinen – beispielsweise aus linguistischer Sicht – nicht belegen lassen und somit möglicherweise durch die Pressegeschichte etwas willkürlich „konstruiert“ wurden. Schließlich wird noch darauf hingewiesen, dass möglicherweise auch der Blick über das 19. Jahrhundert hinaus für die Digitalisierung von Zeitungen von Belang sein könnte, etwa im Hinblick auf das Beispiel der *Frankfurter Zeitung*, die bis 1944 erschien und somit als Untersuchungsgegenstand für interessante historische Anschlussfragen dienen könnte.

Eine Präsentation zu Novellensammlungen des 19. Jahrhunderts (Katharina Herget/Anastasia Pupynina) – wie etwa dem *Deutschen Novellenschatz* von Paul Heyse – und die daran anknüpfende

Diskussion weiten den Blick von Zeitungen auf andere Textarten aus, d.h. vor allem auf Novellen(sammlungen), Zeitschriften und Briefwechsel, deren digitale Verfügbarkeit besonders im Hinblick auf literaturwissenschaftliche Analysen erstrebenswert erscheint. So kann etwa der Novellenschatz als wichtige Station bei der Kanonisierung von Novellen – gleichsam als „Durchlauferhitzer" zwischen Zeitschriften und Werkausgaben – angesehen werden; eine großflächige Digitalisierung bietet also z.B. die Möglichkeit von computergestützten Analysen zu Motiven, Gattungen und der Kanonisierung von Novellen im 19. Jahrhundert. Ebenso wird auf das Interesse an „Flaggzeitschriften" wie der *Deutschen Rundschau* und der *Gartenlaube* in der gegenwärtigen literaturwissenschaftlichen Forschung (z.B. Jannidis, Bayreuth) und entsprechende Anstrengungen bei der Digitalisierung solcher Zeitschriften hingewiesen. Im Lauf der Diskussion wird zunehmend deutlich, dass aufgrund der Vielfalt an möglichen Forschungsfragen aus unterschiedlichen Disziplinen (v.a. Linguistik und Literaturwissenschaft) die Liste von möglicherweise digitalisierbaren Zeitungen, Zeitschriften und anderen Textformen theoretisch „unendlich" fortsetzbar wäre. Somit stellt sich die Frage – zugleich eine zentrale Frage des Workshops –, wie die Liste der potenziell zu digitalisierenden Objekte eingegrenzt werden kann bzw. welche Kriterien hierfür eine Rolle spielen.

II. Kriterien für die Digitalisierung

Es wird noch einmal auf die Kriterien der Überregionalität und Kontinuität von Zeitungen oder Zeitschriften (wie der *Gartenlaube*) hingewiesen, es werden aber auch andere Kriterien ins Spiel gebracht, wie z.B. ästhetische Kriterien, etwa wenn eine Zeitschrift maßgebliche Impulse bei der ästhetischen Gestaltung von Zeitschriften im 19. Jahrhundert gesetzt hat; als Beispiel wird hier die Leipziger *Illustrirte Zeitung* angeführt. Andererseits wird auch auf die Spezifität von überregionalen Zeitschriften hingewiesen, die trotz ihrer allgemeinen Bedeutung oft ein sehr bestimmtes Leserinteresse bzw. -publikum bedienen, wie im Fall der *Gartenlaube*. Als relevantes Kriterium könnte darüber hinaus auch das Vorhandensein einer globalen Leserschaft, wie etwa im Fall der Zeitung *Das Ausland*, gesehen werden. Als für Analysen wichtige Merkmale und Techniken insbesondere von Zeitschriften werden in jedem Fall die Nachbarschaft unterschiedlicher Textgattungen in Zeitschriften, ihre Serialität sowie ihre Wahrnehmungssteuerung genannt. Im Hinblick auf die unterschiedlichen Textgattungen innerhalb von Zeitschriften und Zeitungen wird auch auf Textgattungen hingewiesen, die sich schwer kategorisieren lassen. Auch die Grenze zwischen Zeitungen und Zeitschriften als solchen kann – vor allem in früheren Zeiten – als fließend betrachtet werden. Zudem könnte man eventuell auch die Beilagen von Zeitungen in Betracht ziehen.

Im Folgenden werden einige grundsätzliche Festlegungen der Diskussion kritisch hinterfragt. Die anfangs festgelegte Fokussierung auf den binnendeutschen Raum bei der Auswahl von Textobjekten

wird teilweise kritisiert, da etwa Zeitungen wie die *Allgemeine Zeitung* (v.a. in Augsburg erschienen) durchaus in den österreichischen und osteuropäischen Raum ausgestrahlt haben; über eine generelle Miteinbeziehung dieser Regionen (auch etwa der ehemals deutschen Regionen in Osteuropa) kann somit bei der Auswahl von zu digitalisierenden Objekten nachgedacht werden. Ebenfalls wird die Katalogisierung von (digitalisierten) Zeitungen in Bibliothekskatalogen, vor allem im Hinblick auf die damit verbundenen Zeitangaben, kritisch beleuchtet: So sind Jahrgänge möglicherweise unvollständig oder inkorrekt nummeriert bzw. katalogisiert; die in Bibliothekskatalogen angegebenen Daten zu Zeitungen und ihren digitalisierten Beständen müssen somit eventuell vorsichtig betrachtet werden. Zusätzlich wird angemerkt, dass – gerade im Hinblick auf die Kontinuität von Zeitungen – beispielsweise Besitzer- oder Titelwechsel von Zeitungen die kontinuierliche Betrachtung von Zeitungen erschweren können. Wiederum in Bezug auf eine mögliche Ausweitung des regionalen Raums bei der Betrachtung von Zeitungen wird auf die potenzielle Relevanz deutschsprachiger Textobjekte aus dem nichtdeutschen Raum hingewiesen: So kann die Einbeziehung z.B. von schweizerischen Zeitungen eine vergleichende (kulturwissenschaftliche) Analyse etwa von national-kulturellen Diskursen im Medium der Zeitung erlauben, oder auch (linguistische) Analysen zum Umgang mit Dialekt in Zeitungen. Als Abschluss der Diskussion um die regionale Eingrenzung zumindest von Zeitungen wird eine „pragmatische“ Herangehensweise besprochen, die sich zunächst einmal mit ungefähr zwei Zeitungen aus dem binnendeutschen Raum, genauer: aus dem Raum um Darmstadt/Frankfurt, befassen könnte.

III. Technische Aspekte der Digitalisierung von (historischen) Texten

In einem exemplarischen Blick auf die Digitalisierung einer Zeitung wird aus Erfahrungen im Zuge der Digitalisierung des *Hamburgischen Unpartheyischen Correspondenten* berichtet (Britt-Marie Schuster, Manuel Wille). Hierbei wurden in einem ersten Schritt auf der Grundlage von Scans der Zeitung die Texte per Double Keying-Verfahren transkribiert. Die automatische Erfassung der Texte mittels OCR-Software erwies sich als unzuverlässig, da OCR stark von der Qualität der Scans abhängt und auch mit Frakturschrift (noch) nicht zuverlässig umgehen kann. Die Qualität der in China transkribierten Texte erwies sich als bedeutend besser, wobei hier zudem Layout- und typografische Merkmale der Texte bis zu einem gewissen Grad mitannotiert wurden. Bestimmte Textteile in Zeitungen (etwa Tabellen oder Grafiken) sind jedoch im Allgemeinen schwer transkribier- bzw. annotierbar. Bei den weiteren Digitalisierungsschritten wurde ein TEI/XML-Format des Deutschen Textarchivs (DTA) benutzt, um die Texte in eine im DTA publizierbare Form zu bringen – bislang sind 212 Ausgaben digitalisiert. Dabei wurden weitere, „tiefere“ Annotationen vorgenommen: Etwa das Taggen von Überschriften, Daten, Autoren oder Unterzeichnern von Artikeln, und Orten. Zudem

wurden die Texte durch Tokenisierung, Lemmatisierung und POS-Tagging korpuslinguistisch weitergehend annotiert, sodass sie linguistisch durchsuchbar sind.

Im Zusammenhang mit der Frage nach der Tiefe der Annotation kommt die wichtige Frage auf, wie ausführlich die Annotation im Rahmen einer Digitalisierung notwendigerweise sein sollte. Im Fall des DTA werden etwa unterschiedliche Formate von Texten angeboten, die von reinem Text bis zu tief annotierten XML-Dateien reichen; ersteres wäre z.B. für einige Anwendungsfälle in der Literaturwissenschaft (Topic Modeling) bereits ausreichend, letzteres ist für linguistische Korpusanalysen notwendig. Zentral ist somit die Fragestellung, welchen Zwecken die Digitalisierung und die damit einhergehende Annotation dienen soll; die Texte müssen dann im Hinblick auf diese Zwecke annotiert werden. Mögliche Herangehensweisen an die Digitalisierung können nun darin bestehen, das Korpus von vornherein im Hinblick auf eine bestimmte Forschungsfrage aufzubereiten und zu annotieren, oder aber eine „minimale“ Aufbereitung zu vollziehen und die projektspezifischen Annotationen bzw. Nachbearbeitungen den jeweiligen Projekten zu überlassen. Dabei gibt es durchaus standardisierte Annotationsschritte, die grundsätzlich geleistet würden und auf Konsens und Aushandlungsprozessen innerhalb der (geistes-)wissenschaftlichen Community beruhen; so ist etwa POS-Tagging in vielen Digitalisierungsprojekten Standard. Insgesamt lassen sich solche Digitalisierungsvorhaben unterscheiden, die eher infrastrukturell orientiert sind und bei denen die Digitalisierung der Texte ein zentraler Aspekt ist, und andere Vorhaben, bei denen forschungsorientiert vor allem im Hinblick auf spezifische Forschungsfragen digitalisiert wird und die Digitalisierung somit nur *einen* Schritt innerhalb eines größeren Forschungsvorhabens darstellt. Gerade bei letzteren Projekten lassen sich die Forschungsfragen oder -hypothesen zum Teil durchaus linguistisch formulieren, indem festgestellt wird, an welchen linguistischen Einheiten des Textes sich die Forschungsfragen festmachen lassen, sodass daraufhin eine Annotation der relevanten Einheiten durchgeführt werden kann. Oft handelt es sich dabei jedoch auch um iterative Prozesse, bei denen die Annotation und die Definition/Spezifikation von Forschungsfragen sich wechselseitig ergänzen; je tiefer allerdings annotiert wird, desto idiosynkratischer werden die Annotationen. Andererseits wird auch darauf hingewiesen, dass bei bestimmten, vor allem quantitativ beantwortbaren Forschungsfragen bereits stark von relativ einfach zu bewerkstelligenden Annotationen profitiert werden kann, etwa im Hinblick auf quantitative Konzepte wie Satzlängen oder Worthäufigkeiten. Bei älteren Texten hingegen können solche vermeintlich einfachen Annotation wiederum eine Herausforderung darstellen und einen eigentlichen Teil des Forschungsprozesses darstellen. Art und Umfang der Annotation und Auswahl der entsprechenden Tools hängen jedenfalls auch von Diskursen innerhalb der wissenschaftlichen Community ab – diskutiert werden hierbei etwa Annotationstools wie der *TreeTagger* für Lemmatisierung und POS-Tagging und der *RFTagger* für das Taggen von Kasus. Insgesamt wird festgehalten, dass durch die digitale Aufbereitung der Scans

und des transkribierten Volltextes und durch Basisannotationen wie *TreeTagging* und Satzsegmentierung bereits viel gewonnen wäre: Dies würde einen „common ground“ darstellen, der für viele Projekte und Forschungsfragen als Ausgangsbasis dienen und quantitative Analysen ermöglichen könnte. Es wird auch kurz die Frage nach der Lizenzierung von digitalem Bild- und Textmaterial angesprochen: Bibliotheken etwa scheinen oft bereit zu sein, Scans herauszugeben (wenn sie einmal vorhanden sind); die Volltexttranskriptionen und -annotationen können – wie im Fall des DTA – beispielsweise per *Creative Commons*-Lizenz publiziert werden.

IV. Zur konkreten Realisierung eines Digitalisierungsprojekts

Im Folgenden werden konkrete Arbeitsschritte beim Aufbau eines digitalen Korpus besprochen. So wird ein koordiniertes Vorgehen bei der Digitalisierung vorgeschlagen, bei dem die Arbeit derart aufgeteilt werden könnte, dass zwei Zeitungen mit jeweils unterschiedlichen Forschungszielen digitalisiert werden, um den unterschiedlichen Forschungsinteressen der Teilnehmer gerecht zu werden. Dabei muss zunächst einmal der Arbeitsaufwand für die Digitalisierung zweier Zeitungen ermittelt werden. Die primäre Digitalisierung insbesondere von Zeitungen bietet sich an, da Zeitungen in jedem Fall digitalisiert werden sollen; andere Textobjekte, wie z.B. Zeitschriften, könnten dann zusätzlich digitalisiert werden. Für etwaige Antragstellungen stellt sich wiederum die Frage, ob der Aspekt der Digitalisierung ein Hauptpunkt wäre oder eher Teilaspekt innerhalb eines Projekts, das sich mit weiterführenden Forschungsfragen beschäftigt. Hierbei wurde die Frage nach Projektpartnern und Modi der Zusammenarbeit diskutiert. In diesem Zusammenhang wurde u.a. auf die Infrastruktur und Expertise im Hinblick auf Digitalisierungen im Digital Humanities-Team der TU Darmstadt verwiesen. Die Frage nach der Wichtigkeit des Digitalisierungsaspekts kann pragmatisch durch die Aufteilung in verschiedene Arbeitsschritte angegangen werden: So könnte man zunächst einmal anfangen mit der Digitalisierung unter Nutzung der bereits vorhandenen Ressourcen (in Darmstadt) und in einem zweiten Schritt dann möglicherweise in eher forschungsorientierte Förderprojekte einsteigen. Die literaturwissenschaftliche Gruppe (Thomas Weitin) schlägt außerdem vor, sich später mit Zeitschriftenforschern im Hinblick auf mögliche Digitalisierungsvorhaben in Verbindung zu setzen. Für die Paderborner Gruppe (Britt-Marie Schuster) erscheint vor allem die *Vossische Zeitung* als Untersuchungs- und Digitalisierungsobjekt interessant, wobei im Falle dieser Zeitung noch der genaue Digitalisierungsstand erforscht werden soll; es wird berichtet, dass ein Antrag beispielsweise im Frühjahr 2017 eingereicht werden könnte. Die Digitale Linguistik (Marcus Müller) will sich mit der Digitalisierung der *Frankfurter Zeitung* befassen, wobei auch die infrastrukturelle Machbarkeit der Digitalisierung insgesamt ermittelt werden soll. Ein Anschlusstreffen (z.B. im Frühjahr 2017) wird vorgeschlagen.

V. Personen

Veranstalter:

Prof. Dr. Marcus Müller, TU Darmstadt
Prof. Dr. Thomas Weitin, TU Darmstadt

Teilnehmerinnen und Teilnehmer

Dr. Sabine Bartsch, TU Darmstadt
Dr. Michael Bender, TU Darmstadt
Constanze Hahn, TU Darmstadt
Katharina Herget, TU Darmstadt
David Glück, TU Darmstadt
Dr. Michael Neumann, Universität Konstanz
Anastasia Pupynina, TU Darmstadt
Prof. Dr. Britt-Marie Schuster, Universität Paderborn
Dr. Jörn Stegmeier, Universität Heidelberg
Dr. Tina Theobald, Universität Heidelberg
Manuel Wille, Universität Paderborn