Distant Letters: Methoden und Praktiken zur quantitativen Analyse digitaler Briefeditionen

Dumont, Stefan
dumont@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Haaf, Susanne
haaf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Henny-Krahmer, Ulrike
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Deutschland

Krautter, Benjamin
benjamin.krautter@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Neuber, Frederike
neuber@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Beschreibung

Briefeditionen sind ein Typus der digitalen Edition, in dem die Vorteile des digitalen Mediums bereits mit am intensivsten fruchtbar gemacht werden.[1] In der alltäglichen Arbeit des Edierens sowie der Software-Entwicklung richtet sich der Blick zum großen Teil meist auf den einzelnen Brief und seine Tiefenerschließung, weniger auf die Menge an Briefen eines Korrespondenzkorpus. Weiterführende quantitative Analysen auf Basis der Tiefenerschließung (vollständige Transkription, Modellierung in XML/TEI, Normdaten etc.) und mit digitalen Methoden, die gerade auch für korpusübergreifende Untersuchungen den Weg ebnen würden, sind traditionellerweise in Editionsprojekten (noch) nicht vorgesehen.[2] Mit dem eintägigen Workshop „Distant Letters“ möchten wir ein Panorama an quantitativ orientierten Analysemethoden und -praktiken für Daten digitaler Briefeditionen vorstellen, vermitteln und diskutieren, um so neue Perspektiven auf Briefkorpora zu erproben.[3] Der Workshop gliedert sich in vier Abschnitte:

Auswertung von Metadaten und Entitäten

Auf der Grundlage von standardisiert kodierten Briefmetadaten in XML/TEI sollen mit der Abfragesprache XQuery zunächst Fragen formuliert werden wie: „Wie viel hat Sender A an Empfänger B insgesamt geschrieben? Wie viel in einem bestimmten Jahr?“ Im Anschluss sollen vergleichende Untersuchungen angestellt werden, denen Fragen wie „Wie viel hat Sender A an Empfänger B und Empfänger C geschrieben?“ oder „Wie gestaltet sich das Verhältnis von gesendeten und empfangenen Briefen in der Korrespondenz von A und B?“ zugrunde liegen. Auch Entitäten aus dem Brieftext können in die Untersuchung mit einbezogen werden („Wie häufig wird Person X im Verlauf der Korrespondenz erwähnt?“). Das Ergebnis von derlei Fragen sind statistische Werte, die, um sie interpretatorisch zugänglicher zu machen, weiter aufbereitet werden müssen, z.B. als Visualisierungen in Diagrammen, Kreisen und Kurven.

Analyse linguistischer Merkmale

Im zweiten Teil wendet sich die Untersuchung den Volltextdaten zu. In den Blick genommen werden dabei linguistische Merkmale auf Token-Ebene (z.B. Lemma und Wortart), die einfachen oder komplexen Abfragen (z.B. nach typischen Adjektiv-Anbindungen bestimmter Substantive, Häufungen einer bestimmten Wortart, festen Wendungen, Kollokationen) an den Text zugrunde gelegt werden können und so u.a. Aufschluss über inhaltliche und stilistische Gegebenheiten ermöglichen. Im Workshop werden Werkzeuge gezeigt und benutzt, die zum einen die automatische linguistische Analyse von Texten, z.B. deren Lemmatisierung und POS-Tagging, erlauben und zum anderen Möglichkeiten der Auswertung annotierter linguistischer Merkmale bieten, z.B. mittels leistungsstarker Suchanfragesprachen oder Möglichkeiten der Visualisierung. Genauer in den Blick genommen und z.T. benutzt werden TXM, Corpus Workbench, DTA und WebLicht.[4]

Topic Modeling

Im dritten Teil des Workshops rücken die Inhalte der Briefkommunikation stärker in das Zentrum des Interesses, wenn Fragen aufgegriffen werden wie „Welche Themen werden behandelt und wie sind diese zeitlich verteilt?“ oder „Gibt es bestimmte Themen, die in einer bestimmten Personengruppe stärker verhandelt werden als in einer anderen?“. Analysiert wird dabei der Volltext der Briefe, zusätzlich können jedoch auch die Briefmetadaten in die Interpretation der Analyseergebnisse einfließen. Für die Modellierung der Topics wird das

Tool „Mallet“ verwendet,[5] und es wird im Workshop gemeinsam ein Topic Model für ein Briefkorpus erstellt. Für die Auswertung in Kombination mit Metadaten und Visualisierungen wird der „Topic Modeling Workflow“ (TMW) verwendet.[6] Diskutiert werden soll außerdem, wie sich die Konzepte ‚Topic‘ und ‚Thema‘ zueinander verhalten.[7]

Stilometrie

Im letzten Teil des Workshops soll mit Methoden und Tools der Stilometrie der Sprach- bzw. Schreibstil eines Briefkorpus genauer untersucht werden. Analysiert wird dabei erneut der Volltext, diesmal in orthografisch normalisierter Form. Mögliche Fragestellungen der Analyse sind: „Welche Rückschlüsse erlauben stilometrische Analysen hinsichtlich Sender und Empfänger der Briefe?“, „Korrelieren die stilistische Nähe bzw. Distanz mit Faktoren wie Zeit, Raum oder Empfänger?“. Auch Stilvergleiche werden beispielhaft auf Grundlage der Fragen „Ändert sich der Stil von Sender A in seinen Briefen an die Empfänger B und C?“ und „Variiert der Stil zwischen Geschäfts- und Privatkorrespondenz?“ unternommen. Für die stilometrischen Analysen nutzen wir das „Stylo“-Paket für R.[8] Auch für die stilistischen Analysen ist zu diskutieren, welches Konzept von Stil hinter den gewählten Methoden steht und wie es sich zu anderen Definitionen von Stil verhält.[9]

Ziele

Ziel des Workshops ist es, ein Panorama quantitativer Analysemöglichkeiten für Briefkorpora vorzustellen, das eine Ergänzung zu den traditionellen ‚close reading‘-Verfahren in wissenschaftlichen Editionen darstellt und die Digitalität der Editionsdaten mit Methoden der Digital Humanities noch stärker für quellenimmanente Forschungsfragen fruchtbar macht. Die Teilnehmerinnen und Teilnehmer sollen den Workshop am Ende des Tages mit einem Set an Skripten und Tools verlassen und in der Lage sein, diese auf andere (ggf. eigene) Datensätze anzuwenden. Neben der Vermittlung von technischen Fertigkeiten ist die Diskussion der Methoden und Ergebnisse mit den Teilnehmerinnen und Teilnehmern fester Bestandteil des Workshops. Es soll dabei gemeinsam eruiert werden, auf welchen theoretischen Annahmen die Methoden jeweils basieren, wo ihre Stärken und Schwächen liegen und auch inwieweit die vermittelten Praktiken eine Chance haben könnten, zukünftig ein Bestandteil bei der Erstellung und Nutzung wissenschaftlicher digitaler Briefeditionen zu werden.

Daten

Die Organisatorinnen und Organisatoren stellen XML/TEI und Plain Text Datensätze aus zwei verschiedenen Briefeditionen für den Workshop bereit: ca. 5500 Brieftexte und ebenso viele Metadatensätze aus „Jean Paul - Sämtliche Briefe digital“ (Bernauer / Miller / Neuber 2018) sowie ca. 400 Brieftexte und 3000 Metadatensätze der „edition humboldt digital“ (Ette 2017-). Darüber hinaus steht es den Teilnehmerinnen und Teilnehmer frei, ihre eigenen Datensets (XML/TEI-kodiert und Plain Text) zu verwenden.

Teilnehmerzahl und Vorkenntnisse

Die Anzahl der Teilnehmerinnen und Teilnehmer ist auf 25 begrenzt. Gewisse Grundkenntnisse in der Programmierung (z.B. XSLT/XQuery, Python, R) sind von Vorteil, die im Workshop verwendeten Skripte werden jedoch so vorbereitet, dass sich die Arbeit daran auf Modifikationen und Erweiterungen unter Anleitung der Lehrenden beschränkt. Im Vorfeld des Workshops werden Installationshinweise für die verwendeten Werkzeuge gegeben und die Übungsdaten zum Download bereitgestellt.

Lehrende

Stefan Dumont: Wissenschaftlicher Mitarbeiter bei der TELOTA-Initiative der Berlin-Brandenburgischen Akademie der Wissenschaften, dort u.a. zuständig für die „edition humboldt digital”. Wissenschaftlicher Koordinator des DFG-Projekts „correspSearch - Briefeditionen vernetzen“. Co-Convener der TEI Special Interest Group „Correspondence“. Expertise u.a. mit Standardisierung von Briefkodierung und -metadaten und X-Technologien.

Susanne Haaf: Wissenschaftliche Mitarbeiterin im Projekt CLARIN-D an der Berlin-Brandenburgischen Akademie der Wissenschaften, u.a. beteiligt am Auf- und Ausbau des Deutschen Textarchivs. Doktorandin im Bereich korpusbasierter Untersuchung von Textsortenspezifika. Spezialisierung in Korpusaufbau, Korpuslinguistik, Standards für Text- und Metadaten (insbes. TEI) sowie Textedition.

Ulrike Henny-Krahmer: Wissenschaftliche Mitarbeiterin im Projekt „Computergestützte Literarische Gattungsstilistik” (CLiGS) an der Universität Würzburg. Studium der Regionalwissenschaften Lateinamerika in Köln und Lissabon, Doktorandin in Digital Humanities mit dem Thema „Topic and Style in Subgenres of the Spanish American Novel (1830-1910)“.

Benjamin Krautter: Wissenschaftlicher Mitarbeiter im Projekt “Quantitative Drama Analytics” (QuaDramA) an der Universität Stuttgart. Studium

der Literaturwissenschaft (Germanistik) und Politikwissenschaft in Stuttgart und Seoul (Südkorea), Doktorand im Bereich Digital Literary Studies mit dem Thema “Quantitative Dramenanalyse - Operationalisierung aristotelischer Kategorien” (Arbeitstitel).

Frederike Neuber: Wissenschaftliche Mitarbeiterin bei der TELOTA-Initiative der Berlin-Brandenburgischen Akademie der Wissenschaften, dort u.a. zuständig für die Briefedition “Jean Paul - Sämtliche Briefe digital”. Studium der Italianistik und Editionswissenschaft in Berlin und Rom, Doktorandin in Digital Humanities. Spezialisierung in Editionsphilologie, Datenmodellierung und Programmierung mit X-Technologien.

Fußnoten

1. Der webservice „correspSearch“ etwa illustriert die Bedeutung von standardisierter Metadatenerfassung mit Normdaten zur Vernetzbarkeit von Korrespondenzen, https://correspsearch.net/ .
2. Vereinzelt werden quantitative Analysemethoden bereits auf Editionsdaten angewandt: Etwa wird im Kontext des Projekts “Mapping the Republic of Letters” (Stanford University 2013) zur Erschließung der Briefkommunikation und -verbreitung mit verschiedenen statistisch- und/oder netzwerkanalytisch-basierten Visualisierungen experimentiert; Andorfer (2017) erprobt Topic Modelling mit dem Korrespondenzkorpus Leo von Thun-Hohensteins.
3. Nicht Teil dieses Panoramas ist die Netzwerkanalyse, auch wenn diese Form der Auswertung bzw. Visualisierung für Briefdatensätze oft die am naheliegendste scheint. Grundkompetenzen zur Netzwerkanalyse bzw. -visualisierung werden mittlerweile regelmäßig in Workshops vermittelt, z.B. im Rahmen der „Historical Network Research-Community“ (http://historicalnetworkresearch.org/). Der Fokus des Workshops richtet sich daher auf bisher weniger berücksichtigte Formen der Analyse von Briefkorpora.
4. http://textometrie.ens-lyon.fr , http://www.deutschestextarchiv.de/ , https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page , http://cwb.sourceforge.net/cqpweb.php .
5. http://mallet.cs.umass.edu/topics.php
6. https://github.com/cligs/tmw
7. Zwar ist das Verfahren für die Ermittlung von Schlüsselwörtern und Themen entwickelt worden, je nach verwendetem Korpus ergeben sich aber auch andere Arten von Topics, z.B. sprachspezifische oder motivische. Vgl. dazu u.a. Rhody (2012) und Schöch (2017).
8. https://sites.google.com/site/computationalstylistics/stylo
9. Für einen Überblick zu verschiedenen Stilbegriffen in der Literatur- und Sprachwissenschaft siehe Herrmann et al. (2015).

Bibliographie

Andorfer, Peter (2017): *Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich*, in: Zeitschrift für digitale Geisteswissenschaften; doi: 10.17175/2017_002 [zuletzt abgerufen 7. Januar 2019].

Bernauer, Markus / Miller, Norbert / Neuber, Frederike (eds.) (2018): *Jean Paul – Sämtliche Briefe digital.* In der Fassung der von Eduard Berend herausgegebenen 3. Abteilung der Historisch-kritischen Ausgabe (1952-1964), im Auftrag der Berlin-Brandenburgischen Akademie der Wissenschaften überarbeitet und herausgegeben von Markus Bernauer, Norbert Miller und Frederike Neuber; http://jeanpaul-edition.de [letzter Zugriff 7. Januar 2019].

Burrows, John (2002): *Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship*, in: Literary and Linguistic Computing 17/3, S. 267–287.

Dumont, Stefan (2016): *correspSearch – Connecting Scholarly Editions of Letters*, in: Journal of the Text Encoding Initiative [Online], Issue 10; http://journals.openedition.org/jtei/1742 [letzter Zugriff 7.Januar 2019].

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): *Stylometry with R: A Package for Computational Text Analysis*, in: The R Journal 8/1 (2016), S. 107–121.

Ette Ottmar (eds.) (seit 2016): *edition humboldt digital.* Berlin-Brandenburgische Akademie der Wissenschaften, Berlin. Version 3 vom 14.09.2018; https://edition-humboldt.de/ [letzter Zugriff 7. Januar 2019].

Graham, Shawn / Weingart, Scott / Milligan, Ian (2012): *Getting Started with Topic Modeling and MALLET*, in: The Programming Historian 1; https://programminghistorian.org/en/lessons/topic-modeling-and-mallet [letzter Zugriff 7. Januar 2019].

Heiden, Serge (2010): *The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme*, 24th Pacific Asia Conference on Language, Information and Computation, Nov 2010, Sendai, Japan. Institute for Digital Enhancement of Cognitive Development, Waseda University, S.389–398.

Herrmann, Berenike J. / van Dalen-Oskam, Karina / Schöch, Schöch (2015): *Revisiting Style, a Key Concept in Literary Studies*, in: Journal of Literary Theory 9/1, S. 25–52.

Rhody, Lisa M. (2012): *Topic Modeling and Figurative Language*, in: Journal of Digital Humanities 2/1; http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ [letzter Zugriff 7. Januar 2019].

Schöch, Christof (2017): *Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama*, in: Digital Humanities Quarterly 11/2; http://www.digitalhumanities.org/dhq/

vol/11/2/000291/000291.html [letzter Zugriff 7. Januar 2019].

Walmsley, Priscilla (2009): *XQuery: Search Across a Variety of XML Data*. O'Reilly Media.