Historische Korrespondenzen und Social Media Analytics. Eine experimentelle Analyse der Briefe aus Jean Pauls Umfeld

Frederike Neuber

Einleitung

Wenn die historische Briefkultur um 1800 Ähnlichkeiten zu den Social Media der Gegenwart aufweist, dann sollten die Strukturen und Praktiken der epistolaren Kommunikation auch mit ähnlichen Methoden untersuchbar sein – so lautet die Eingangsthese dieses Beitrags. Während wissenschaftliche Editionen traditionell Ergebnis und Gegenstand qualitativer Forschung sind, in denen der Einzeltext oder wenige Texte im Zentrum der Beobachtung und Argumentation stehen, werden die umfangreichen Datenkorpora der Social Media vorrangig aus quantitativer Perspektive und mit digitalen Methoden analysiert. Durch den digitalen Wandel der Gesellschaft ist das Sammeln, Überwachen, Analysieren und Visualisieren von Informationen aus den sozialen Medien im letzten Jahrzehnt für die verschiedensten Disziplinen und Arbeitsfelder immer relevanter geworden, darunter Kommunikationswissenschaft, Wirtschaft, Informatik, Politik und öffentliche Verwaltung.[[1]](#footnote-1) Diesem Kontext entstammen die Social Media Analytics, die als interdisziplinäres Forschungsfeld und Querschnittsmethode wissenschaftliche Verfahren sowie Software zur Aggregation und Analyse von sozialen Interaktionen und Inhalten entwickeln.[[2]](#footnote-2) Das damit ermöglichte ‚distant reading‘[[3]](#footnote-3) von Daten, das sich von der Detaillektüre einzelner Texte und deren Inhalten löst und eine abstrakte bzw. übergeordnete Perspektive auf einen größeren Textbestand generiert, kann – übertragen auf historische Korrespondenzen – dazu beitragen, allgemeine Merkmale und Muster der Briefkommunikation zu identifizieren.[[4]](#footnote-4)

Anhand des Korrespondenzumfelds des Jean Pauls, einer ‚Social Media Community‘ um 1800, bestehend aus Familie, Freundinnen und Kolleginnen des Schriftstellers, erprobt der Beitrag die Übertragung von Methoden der Social Media Analytics auf die Analyse historischer Briefkorpora.[[5]](#footnote-5) Seit 2019 wird ein ausgewählter Briefbestand aus Jean Pauls Umfeld an der Berlin-Brandenburgischen Akademie der Wissenschaften ediert und erscheint seit 2020 als Teil der Editionsplattform *Jean Paul – Sämtliche Briefe digital*.[[6]](#footnote-6) Ausgehend von den Gemeinsamkeiten zwischen epistolarer und digitaler Kommunikation wird der Beitrag zunächst einen Vergleich zwischen Social Media-Daten und Editionsdaten ziehen sowie den angewendeten Analyseworkflow illustrierten. Es folgt eine Analyse des Korpus der Umfeldbriefe nach drei Metriken bzw. Kennzahlen der Social Media Analytics, die Auskunft über die Rolle von Korrespondentinnen, die Gewichtung der verhandelten Inhalte sowie die Tonalität der Kommunikation geben: *Reach, Share of Voice* und *Opinion*.[[7]](#footnote-7) Es geht im Wesentlichen darum *wer* kommuniziert, *worüber* kommuniziert wird und *wie* kommuniziert wird, wobei die drei genannten Kennzahlen im Kontext des jeweiligen Analyseabschnitts genauer definiert werden. Der Beitrag schließt mit einer Diskussion der Ergebnisse.

Umfeldbriefe und Social Media

Soziale Netzwerke sind kein reines Internetphänomen. Konstitution und Dynamik digitaler Netzwerke wie Facebook und Instagram weisen einige Ähnlichkeiten zu den Briefnetzwerken des 18. und 19. Jahrhunderts auf. Auf beiden ‚Plattformen‘ finden sich Communities, d. h. Gruppen von Individuen mit gemeinsamen Interessen und Bekanntschaften, zusammen. Das Umfeld Jean Pauls ergibt sich durch die gemeinsame Bekanntschaft zum Schriftsteller, die direkt oder über Dritte bestehen kann. Zum derzeit aus 1156 Briefen bestehenden Editionskorpus (Stand Mai 2022, v.5.0[[8]](#footnote-8)) zählen u. a. Caroline Richters (geb. Mayer, 1777–1860) eigenständige Korrespondenz jenseits der Briefe mit ihrem Ehemann Jean Paul, die Briefe ihrer Schwester und Berufsschriftstellerin Minna Spazier (geb. Mayer, spätere Uthe, 1776–1825) sowie die Korrespondenzen von Jean Pauls Freunden, darunter der vielvernetzte jüdische Kaufmann Emanuel (Osmund, 1766–1842). Innerhalb des Korpus lassen sich verschiedene Kommunikationszirkel ausmachen, wie die Kommunikation des Ehepaars Richter mit den Kindern Emma (1802–1853), Max (1803–1821) und Odilie (1804–1865) oder das Korrespondenznetz, das sich während Jean Pauls Zeit in Weimar bildete, dem u. a. Caroline Richter und Johann Gottfried Herder (1744–1803) angehören.

Der Kitt der epistolaren Kommunikation sind gemeinsame Interessen und Themen. Ähnlich wie Userinnen auf Instagram heutzutage unter den Hashtags #travelgram oder #instatravel von ihren Reisen berichten, ergeben sich im Umfeldnetzwerk thematische Bündelungen der Briefe zum Thema Reisen, wenn auch ohne Markierung durch ein entsprechendes Hashtag. Sowohl in den Social Media des World Wide Web als auch im sozialen Medium Brief gestalten die User die Inhalte selbst (i. e. ‚user generated content‘), weshalb diese „Informationen zu allen möglichen Lebensbereichen enthalten [können], die kaum professionell oder institutionell gefiltert würden oder den Anspruch hätten, Teil einer soziopolitischen Öffentlichkeit zu sein.“[[9]](#footnote-9) Dies gilt auch für die Edition der Umfeldbriefe, die durch die Verhandlung von Themen wie Kindererziehung, Feste und Feierlichkeiten, Gesundheit und Krankheit oder finanzielles Haushalten ein Einblick in das Leben bürgerlicher Kleinfamilien um 1800 gibt. Die Kommunikationsformen sind damals wie heute teilweise sehr stark konventionalisiert, um die Beziehung zu Korrespondenzpartnerinnen oder die Haltung zu bestimmten Themen auszudrücken. Während es in den Social Media rhetorische Marker wie Emojis gibt, mit denen man die Tonalität des Geschriebenen explizit machen kann, ist die ‚Stimmung‘ in historischen Briefen dem Text und seiner sprachlichen Gestaltung inhärent.[[10]](#footnote-10)

Editionsdaten als ‚Smart Data‘

Sowohl bei Social Media-Inhalten als auch bei digitalen Editionstexten handelt es sich im Kern um Datensätze,[[11]](#footnote-11) die sich hinsichtlich einiger Aspekte in ihrer Konstitution unterscheiden. Datensätze sozialer Netzwerke kann man, insofern sie frei verfügbar sind, über technische Schnittstellen oder über spezielle Monitoring-Tools beziehen, wobei, letztere vor allem von Unternehmen zur Auswertung von Interaktionen oder Inhalten verwendet werden, um zukünftige Entwicklungen wie Trends und Kundeninteressen zu prognostizieren. Auch wenn die Daten der Social Media für Auswertungen meist vorverarbeitet werden, ist die Überlieferung selbst Ausgangsdatensatz. Demgegenüber wird die Überlieferung historischer Korrespondenzen durch die digitale Edition bzw. ihre Kodierung repräsentiert, in der editorische Entscheidungen über die Erfassung und Erschließung des Textes zum Tragen kommen. Die Distanz zwischen originärer Kommunikation und Analyse ist bei historischen Korrespondenzen also wesentlich größer als bei Social Media Korpora. Die editorische Bearbeitungsschicht kann den Informationsgehalt der Überlieferung einerseits reduzieren, wenn beispielsweise textkritische Phänomene stillschweigend normalisiert werden, sie andererseits aber auch mit neuem Wissen anreichern, beispielsweise durch die Referenzierung von Personennamen auf Normdatensätze.[[12]](#footnote-12)

Daten aus Social Media-Plattformen und Daten aus wissenschaftlichen Editionen unterscheiden sich im Wesentlichen in Qualität und Umfang. Die „Sozialen Medien [bieten] zwar eine Überfülle von Informationen (Big Data), aber zunächst einmal keine hoch-qualitativen Daten im herkömmlichen Sinn,“[[13]](#footnote-13) wie sie in digitalen Editionen vorliegen und welche in die Kategorie ‚Smart Data‘ fallen, die Christoph Schöch wie folgt definiert:

Smart data is data that is structured or semi-structured; it is explicit and enriched, because in addition to the raw data, it contains markup, annotations and metadata. And smart data is “clean”, in the sense that imperfections of the process of capture or creation have been reduced as much as possible, within the limits of the specific aspect of the original object being represented. This also means that smart data tends to be “small” in volume, because its creation involves human agency and demands time.[[14]](#footnote-14)

Mit 1156 Dokumenten ist das Korpus der Umfeldbriefe sowohl aus Sicht der Social Media Analyse als auch aus Perspektive der quantitativen Textanalyse ein eher kleines, aber ein ‚intelligentes‘ Datenset. Die Briefe liegen im Standardformat XML vor und sind nach den Richtlinien der Text Encoding Initiative (TEI), dem de facto Standard zur Kodierung wissenschaftlicher digitaler Editionen, erfasst.[[15]](#footnote-15) Das Kodierungsschema der Umfeldbriefe orientiert sich an zwei TEI-Subsets: dem Basisformat des Deutschen Textarchivs[[16]](#footnote-16) und ediarum.BASE, welches der Software ediarum,[[17]](#footnote-17) mit dem im Rahmen der Umfeldbriefe ediert wird, zugrunde liegt. Mit der TEI werden sowohl strukturelle Informationen zur Textgestalt (z. B. Überschrift, Absatz, Zeile etc.) als auch inhaltlich angereicherte Informationen (z. B. zu textkritischen Phänomenen, zu Personen- oder Ortsnamen) explizit kodiert. Damit kann man diese Informationen nicht nur in einer digitalen Edition visualisieren, sondern sie auch computergestützt auswerten, um Muster, Beziehungen, Trends, Korrelationen und Anomalien in Bezug auf das Gesamtkorpus zu analysieren.[[18]](#footnote-18)

Entscheidend für ‘smarte Daten’ ist das Vorhandensein und die Qualität von Metadaten, d.h. von Informationen, mit denen das eigentliche Datenset beschrieben wird. Für die im Beitrag vorgenommenen Analysen sind im Korpus der Umfeldbriefe aus dem Bereich des <teiHeader>, dem TEI-Abschnitt zur Verzeichnung der Metadaten, vor allem zwei Informationstypen interessant: Briefmetadaten und Verschlagwortung.

Im TEI-Element <correspSearch> werden die Briefmetadaten strukturiert erfasst, d. h. Informationen zu Sender:innen, Sendedatum und -ort sowie Empfänger:innen (sowie, wenn vorhanden Empfangsdatum und -ort). Die Jean Paul Edition setzt dabei auf die Verwendung von Normdatensätzen. Personen sind, wenn vorhanden, mit Identifikatoren der Gemeinsamen Normdatei[[19]](#footnote-19) verknüpft, Orte mit Identifikatoren von GeoNames.[[20]](#footnote-20) Datumsangaben sind einheitlich im Format der International Organization for Standardization (ISO)[[21]](#footnote-21) erfasst. Der hohe Grad an Normierung der genannten Informationen ist einerseits die Voraussetzung, um einen Bezug zwischen Dokumenten bzw. Entitäten in der Edition herzustellen. Andererseits ermöglichen die Standards eine Einbindung der Metadaten in andere Kontexte, wie beispielsweise in den Webservice correspSearch,[[22]](#footnote-22) der Briefmetadaten verschiedener Editionen aggregiert. Nicht zuletzt stellt die stringente Erfassung der Briefmetadaten die Weichen für kontextualisierbare Analysen, mit denen sich bestimmte Fragen an die Inhalte des Briefkorpus in Bezug zu Personen oder Zeiträumen setzen lassen.

Im Gegensatz zu den Briefmetadaten zählt die in der Edition vorgenommene Verschlagwortung der Dokumente nicht zu den gängigen Metadatenkategorien digitaler Korrespondenzeditionen. Um die Vielstimmigkeit der Community in den Daten abzubilden und um inhaltliche Cluster zu bilden, wurde ein zweigliedriges Schlagwortsystem entwickelt, über das jeder Brief einerseits einem oder mehreren Korrespondenzkreisen zugeordnet wird und andererseits mit thematischen Schlagworten inhaltlich kategorisiert wird. Die Verschlagwortung wird innerhalb des TEI-Abschnitts <textClass> mit Referenz auf Registerdateien, die von den Bearbeiter:innen der Edition gepflegt werden, realisiert.[[23]](#footnote-23) Zwei größere Korrespondenzkreise bzw. Teilcommunities im Umfeld sind beispielsweise der *Caroline Richter-Kreis* (189 Briefe), der die Korrespondenz um Jean Pauls Frau aggregiert, oder der *Zeitung für die elegante Welt-Kreis* (113 Briefe), in dem sich die Korrespondenz um die 1800 von Jean Pauls Schwager Karl Spazier gegründete gleichnamige Zeitschrift sammelt. Zu den etwas kleineren und weniger spezifischen Kreisen zählen beispielsweise die *Verlagskorrespondenz* (72 Briefe) und die *Amtskorrespondenz* (9 Briefe). Zu den thematischen Schlagworten zur systematischen Erfassung der Kommunikationsinhalte zählen u.a. *Reisen* (147), *Berufliches* (17 Briefe) und *Bittschreiben* (15 Briefe). Allgemeine und viel behandelte Themen können im zweistufig angelegten Themenregister weiter ausdifferenziert sein, wie *Reisen* in u. a. *Ausflüge* (7), *Besuche* (117) und *Jean Pauls Besuche in Heidelberg* (7). Die Themenverschlagwortung ist – über die Edition hinaus – ein wertvolles Instrument zur Erforschung des Familienlebens und -alltags um 1800. Themen wie *Erziehung*, *Geschlechterrollen* und *Krankheiten* sind über den literaturwissenschaftlichen Schwerpunkt an der Edition hinaus auch für die Sozial- und Kulturwissenschaften sowie für die Geschlechterforschung von Interesse. Derzeit sind insgesamt 40 Korrespondenzkreise und 64 Themen mit insgesamt 80 Unterthemen verzeichnet.

Auch wenn es sich bei dem Korpus der Umfeldbriefe wie bereits erwähnt nicht um ‚Big Data‘ handelt, ist für die geplanten Analysen ausschlaggebend, dass sich aufgrund des hohen Grades an homogener technischer Strukturierung und informationsreicher inhaltlicher Annotation Erkenntnisse über die kommunizierende Community ableiten lassen. Allgemeiner formulieren es danah boyd und Kate Crawford: “The size of data should fit the research question being asked; in some cases, small is best.”[[24]](#footnote-24)

Analyseverfahren

Stieglitz und Dang-Xuan beschreiben in ihrem *Social Media Analytics Framework[[25]](#footnote-25)* Datenanalyse als fünfstufiges Verfahren, das sich wie folgt zusammensetzt:[[26]](#footnote-26) i.) die Definition einer Zielstellung der Analyse, ii.) die Sammlung von Daten, beispielsweise über technische Schnittstellen oder HTML-Parsing, iii.) die Vorverarbeitung der gesammelten Daten für die Analyse (Englischen „Preprocessing“), bei der beispielsweise irrelevante Informationen aus den Daten entfernt werden, iv.) die eigentliche Datenanalyse und v.) der Ergebnisbericht. Bei der Datenanalyse (iv.) wird zwischen drei Ansätzen differenziert: a.) die Themen der Kommunikation, die vorrangig durch Text Mining-Verfahren ermittelt werden, b.) die Stimmung der Kommunikation, identifiziert im Verfahren der Sentiment Analyse, und c.) die Kommunikationsstrukturen, ausgewertet in Netzwerkanalysen.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

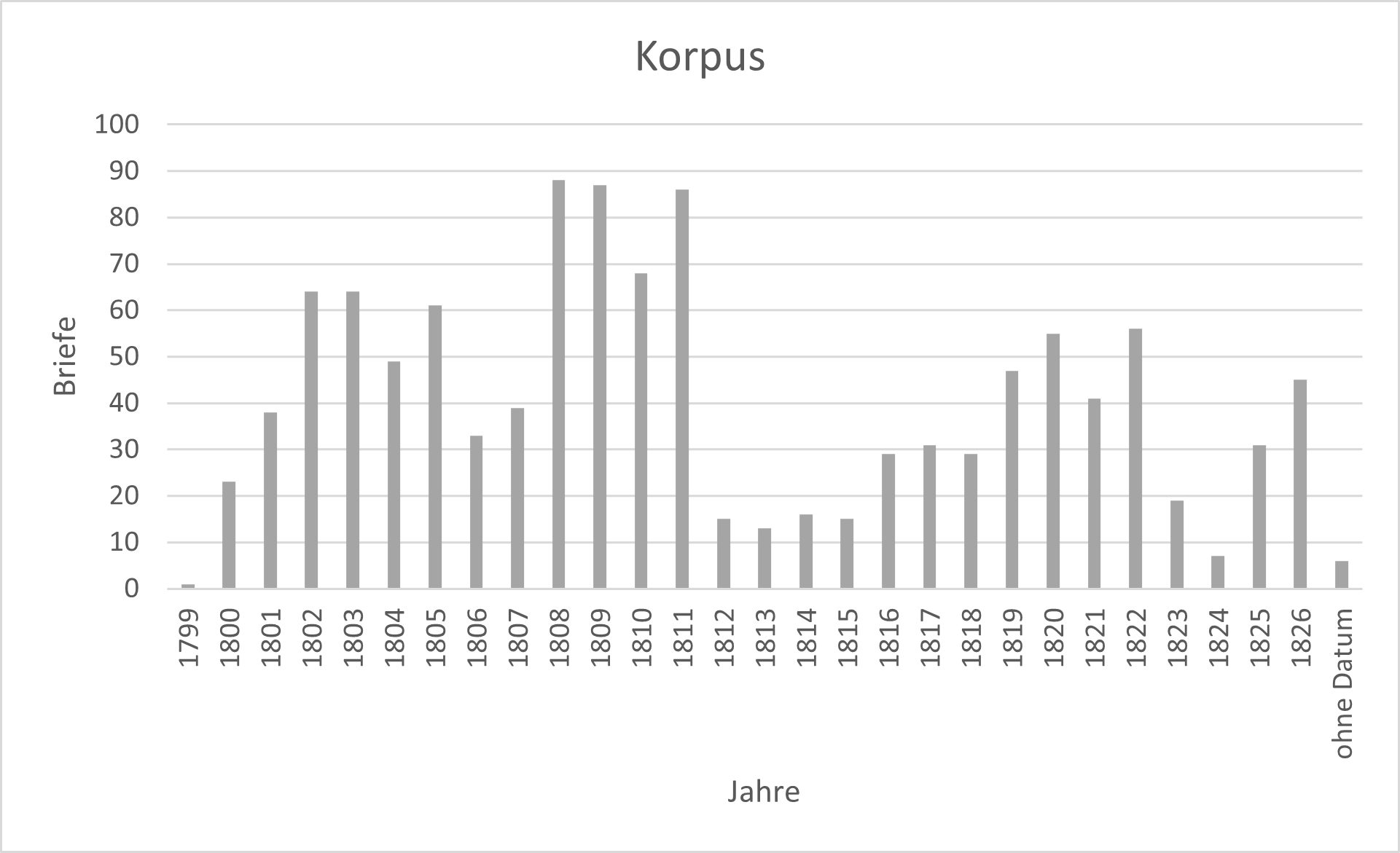
Der Workflow bei der Analyse der Umfeldbriefe entspricht im Wesentlichen dem Modell von Stieglitz und Dang-Xuan (Abb. 1). Die drei Fragenbereiche nach dem „Wer?“, dem „Worüber?“ und dem „Wie?“ der Kommunikation bilden den Ausgangspunkt. Die Daten werden nicht live, sondern über das Forschungsdatenrepositorium zenodo bezogen, wo sie unter Creative Commons-Lizenz verfügbar sind. Diese Ausgangsdaten werden so vorverarbeitet (engl. preprocessing), dass ein Analysedatenset entsteht, das auf die Aspekte reduziert ist, die zur Beantwortung der Fragestellungen relevant ist. Was die eigentliche Analyse angeht, so weicht das Vorgehen vom *Social Media Analytics Framework* in zwei Hinsichten ab: Zum einen schlägt das Modell, bedingt durch den geringen Strukturierungsgrad von Social Media Daten, vorrangig Verfahren des maschinellen Lernens zur Analyse vor, während bei der Analyse der Umfeldbriefe nicht ausschließlich, aber vorranging bereits erfasste Datenfelder (z.B. Schlagworte) statistisch ausgewertet wurden. Die Personen bzw. ihre Rolle im Korpus sowie Untersuchung der verwendeten Themen, wir auf Basis in den Daten vorhandener Felder statistisch ausgewertet. Die Tonalität der Kommunikation wird hingegen mit Sentiment Analysis, einem Verfahren aus dem Natural Language Processing vorgenommen.

Die zweite Abweichung bei der Analyse der Umfeldbriefe zu Stieglitz‘ und Dang-Xuans Modell ist genereller Natur. Während das Social Media Analysemodell nahelegt, die Auswertung der Texte verlaufe linear von der Eingabe eines Datensatzes bis zur Ausgabe der Ergebnisse, muss man sich die ‚Laufrichtung‘ zwischen Preprocessing (iii.), Analyse (iv.) und Ergebnis (v.) vielmehr als Spirale vorstellen, bei der sich mit jeder Drehung das Verständnis über die Analyse selbst erweitert:[[27]](#footnote-27) Zum einen kann es beispielsweise sein, dass man in einem ersten Analyseansatz zu einem Ergebnis gelangt, dass durch den neuen Blick auf die Daten den Impuls gibt, bisher in den Analysedaten nicht berücksichtigte Datenfelder doch in die Analyse aufzunehmen. Zum anderen kann sich das Ergebnis als unbefriedigend herausstellen, wenn beispielsweise zu viele Merkmale bei der Analyse berücksichtigt wurden und das Ergebnis dadurch überfrachtet ist. In solch einem Fall würde man Analysedaten oder Methode ebenfalls entsprechend anpassen, um den Fokus der Analyse stringenter herauszustellen.

Preprocessing

Für die drei vorgenommenen Auswertungen wurden zwei Analysedatensets generiert, die, gemeinsam mit den Ergebnisdaten und deren Visualisierungen auf GitHub verfügbar sind.[[28]](#footnote-28) Zum einen wurde ein *Metadatenkorpus* in XML erstellt, das im Wesentlichen aus editorisch angereicherten Informationen besteht und keine Brieftexte enthält. Die XML-Dokumente umfassen briefweise Titeldaten, Korrespondenzmetadaten (Senderin, Empfängerin, Sendedatum, Sendeort) sowie Schlagworte zu Korrespondenzkreisen und Themen. Zum anderen wurde ein Analysekorpus nur aus *Textdaten* erstellt, und zwar in verschiedenen Arrangements, u. a. sortiert bzw. gruppiert nach Senderinnen, Jahren und Schlagworten. Die Brieftexte wurden dafür im Vorfeld mit Hilfe der Software CAB[[29]](#footnote-29) – kurz für C in historischer Orthografie und Varianz normalisiert, um textbasierte Analysen zu vereinfachen. Editorische Anreicherungen wurden gänzlich entfernt. Für die nachfolgenden Analysen nach Reichweite und Share of Voice wurde das Metadatenset mit XSLT prozessiert und die Ergebnisse als Tabellen in CSV generiert sowie mit MicrosoftExcel in verschiedenen Diagrammen visualisiert. Das Textdatenset wurde mit dem Tool SentText für die Ermittlung der Opinion ausgewertet.[[30]](#footnote-30)

Vor den eigentlichen Analysen noch zwei Hinweise zur Datengrundlage, der für den Umgang und die Deutung der Ergebnisse relevant ist: Zum einen wächst das Korpus, das derzeit in Version 5.0 vorliegt, sukzessive, so dass sich die Ergebnisse bei identischer Analyse bei Version 6.0 oder 7.0 schon deutlich anders gestalten können, wenn sich die Datengrundlage, beispielsweise durch die Integration einer größeren Teilkorrespondenz, ändert. Zum anderen liegt es gewissermaßen in der Natur des Umfelds, keinen klaren Abschluss zu haben, weshalb die Editorinnen diesen durch eine Selektion bzw. den Ausschluss von Quellen selbst bestimmen. Diese Auswahl kann darüber hinaus je nach Überlieferungslage Lücken aufweisen. Außerdem unterliegt sie bereits gewissen Vorannahmen bzw. Biases der Editorinnen, die sich folglich in den Analyseergebnissen widerspiegeln. Nicht nur im Kontext der Auswertung der Umfeldbriefe, sondern für jegliche Datenanalysen ist daher zentral, die Ergebnisse auf die Datengrundlage, nicht aber auf die Wirklichkeit zu beziehen.



**### Abb. 1: Jahrweise Verteilung der Briefe im Korpus. ###**

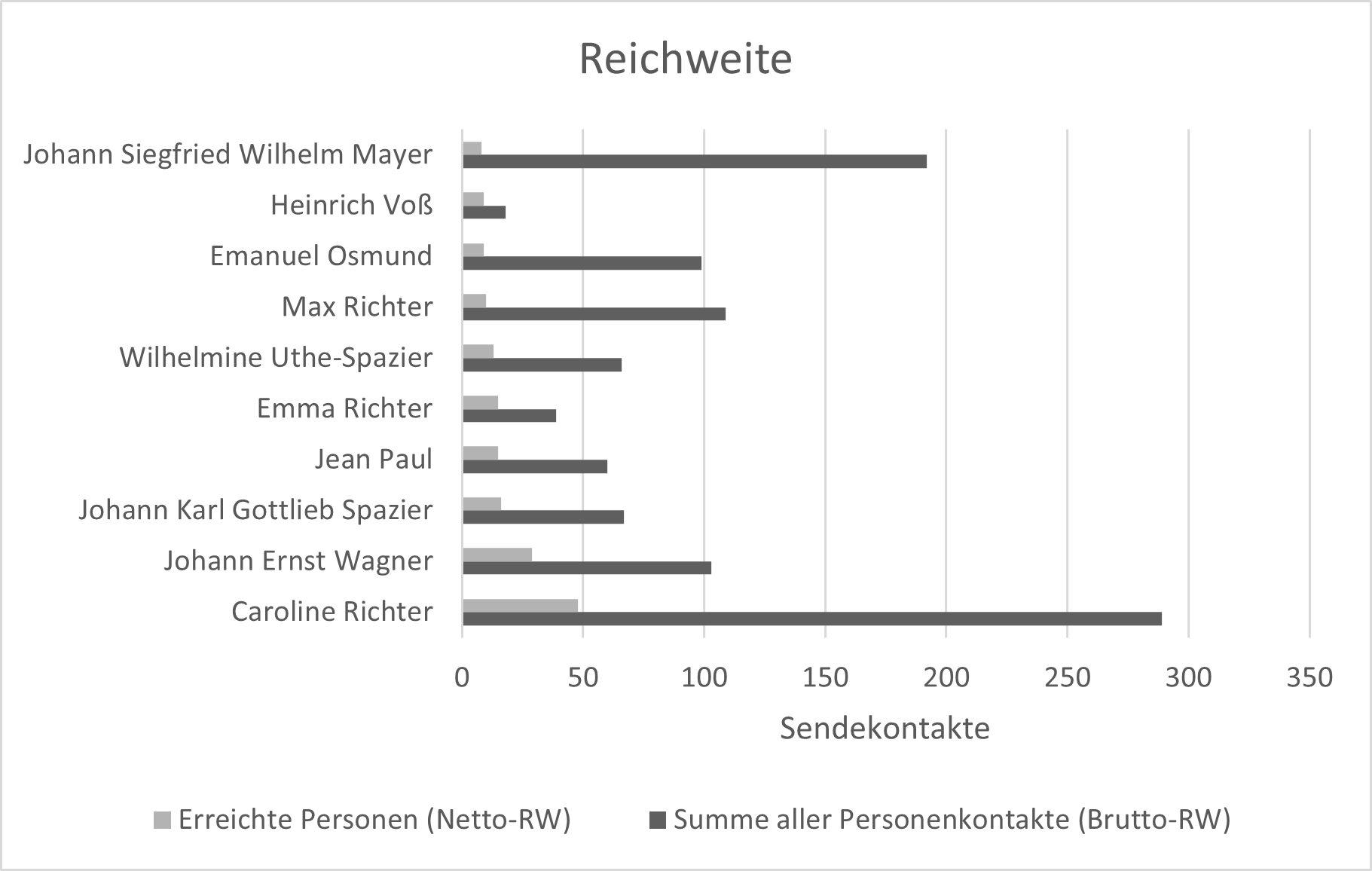
Um die Datengrundlage besser kennenzulernen, empfiehlt sich zu Beginn ein jeder Analyse eine Art ‚Bestandsaufnahme‘, um bestimmte Parameter der Daten zunächst einmal einfach zu zählen. Eine solche Zählung ergab, dass die 1156 Briefen von 109 Verfasserinnen teilweise allein (), teilweise kollaborativ () erstellt wurden. 129 Personen waren explizit Empfängerinnen der Briefe, 26 Personen haben Briefe anderer Empfängerinnen mitgelesen. Zwischen den drei Rollen, Verfasserin, Empfängerin und Mitleserin kann es Überschneidungen geben. Eine jahrweise Zählung der Briefe zeigt (Abb.1), dass ihre Verteilung über das Korpus nicht ausgewogen ist, was man, insofern sich Berechnungen auf Jahre beziehen, berücksichtigen muss.

Reichweite

Allgemein gefasst versteht man unter dem Begriff Reichweite (engl. Reach), wie viele Personen erreicht werden können, und zwar je nach Kontext durch verschiedene Kommunikationsträger wie etwa Websites im Allgemeinen, Werbeträger und Marken, die über Websites präsentiert werden, sowie Personen, meist Influencerinnen.[[31]](#footnote-31) Für das Korpus der Umfeldbriefe soll die Reichweite der Verfasserinnen in den Blick genommen werden, was zunächst einmal erfordert, die Berechnung so anzulegen, dass die multipolaren Kommunikationsstrukturen berücksichtigt werden. Anstatt in Briefen als Bezugsgröße zu rechnen, wird jedes Verfassen eines Briefes – ob allein oder gemeinschaftlich – als eigener ‚Sendekontakt‘ gewertet, mit dem sich die verschiedenen Verfasserinnen sich in das Kommunikationsnetzwerk einbringen und andere Personen erreichen. Bedingt durch das kollaborative Verfassen von Briefen liegt die Summe aller Sendekontakte mit 1247 höher als die Summe der Briefe (1156).[[32]](#footnote-32)

Im Schnitt geht von jeder Korrespondentin rund 11mal ein Sendekontakt aus. Sieht man genauer hin, dann haben von den 109 Verfasserinnen aber lediglich 23 Personen 11 oder mehr Briefe allein oder gemeinschaftlich verfasst, d.h. rund 21%. Allein die 5 aktivsten Verfasserinnen machen mit 792 Sendekontakten rund 64% der Kommunikation im ganzen Korpus aus: Caroline Richter (289 Sendekontakte), Johann Siegfried Wilhelm Mayer (192), Max Richter (109), Johann Ernst Wagner (103) und Emanuel Osmund (99). An dieser Stelle kann mal also zunächst einmal festhalten, dass, was die Anzahl der Sendekontakte bzw. Briefe angeht, im derzeitigen Datenbestand eine enorme Ballung allein fünf Verfasserinnen liegt, wobei Caroline Richters Sendekontakte allein bereits 23% aller Sendekontakte ausmachen.

Die Zählung der Sendekontakte gibt eine erste Übersicht, um die Gewichtungen im Korpus besser zu verstehen, sagt aber nur bedingt etwas über die Reichweite der Verfasserinnen aus. Es bietet sich daher an, in einer weiteren Analyse zwischen der Summe aller Sendekontakte einer Person, bei der Überschneidung von gleichen Empfängerinnen nicht berücksichtigt werden, und der Zahl der verschiedenen Empfängerinnen, die mit dem Sendekontakt erreicht wurden, zu unterscheiden. Im Online-Marketing differenziert man diese beiden Größen als *Bruttoreichweite* und *Nettoreichweite*.[[33]](#footnote-33)



**### Abb. 2: Brutto- und Nettoreichweite der 10 Verfasserinnen mit der höchsten Nettoreichweite ###**

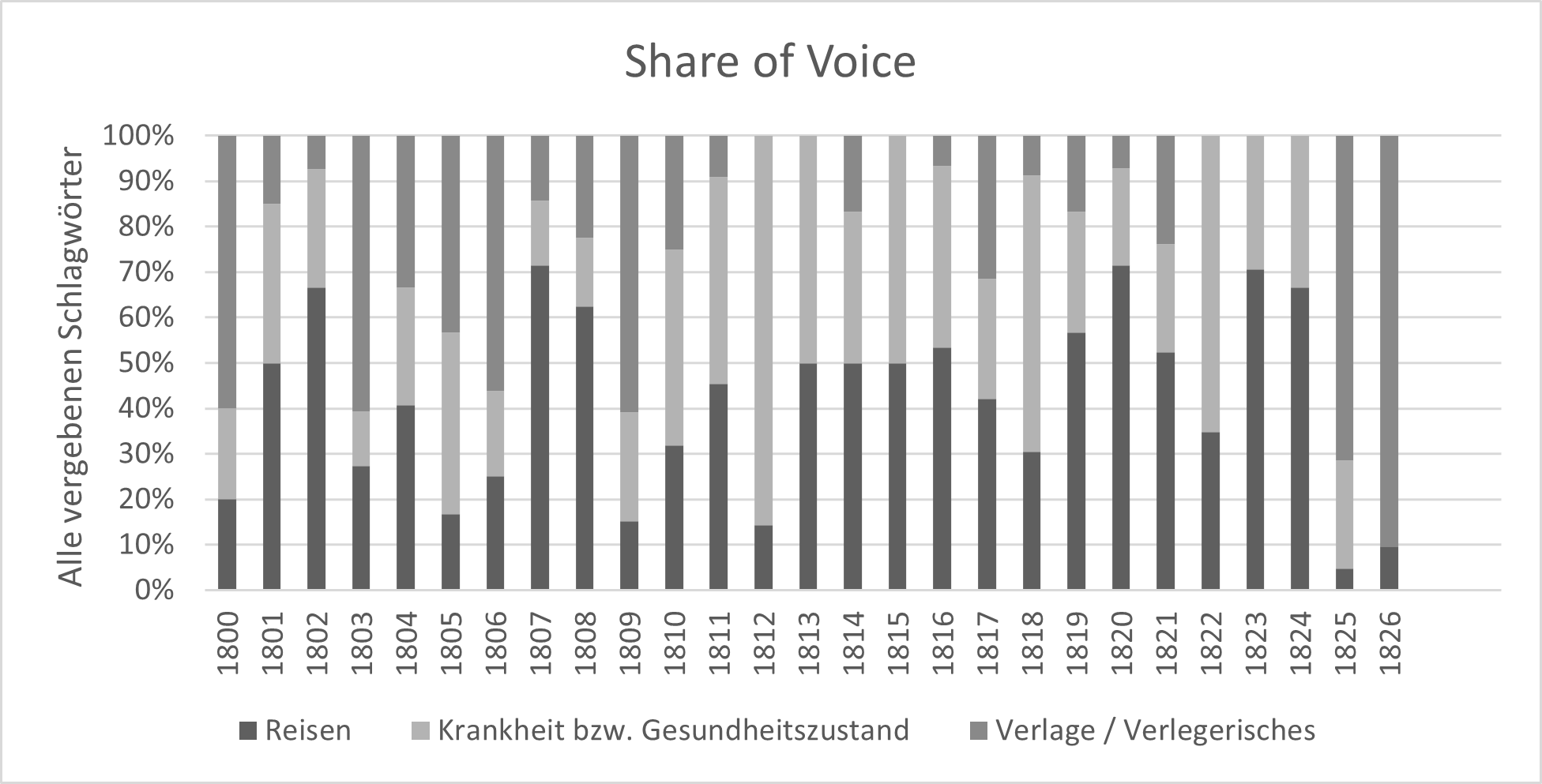
Abbildung 2 zeigt das visualisierte Ergebnis einer solchen Auswertung. In der Visualisierung wurden lediglich die 10 Senderinnen mit der höchsten Nettoreichweite berücksichtigt, also mit den meisten Sendekontakten ohne Überschneidung der Empfängerinnen (jeweils oberer Balken). Gleichzeitig wird dieser Wert in Bezug zur Anzahl aller Sendekontakte (jeweils unterer Balken) gestellt. Caroline Richter, deren hohe Bruttoreichweite bereits angesprochen wurde, hat gleichzeitig die höchste Nettoreichweite. Ihr Vater hingegen, Johann Siegfried Wilhelm Mayer, der anteilig die zweithöchste Summe an Sendekontakten im Korpus hat, hat eine geringe Nettoreichweite, da er hauptsächlich mit seiner Tochter korrespondierte. Johann Ernst Wagner der mit XXX Briefen im der Kalkulation der Bruttoreichweite eher im //Schule Heinrich Voß //hohe Reichweite, warum?

Share of Voice

Bei der Ermittlung und Auswertung von Themen werden im Bereich der Social Media Analytics heutzutage oftmals Verfahren des maschinellen Lernens wie Topic Modeling eingesetzt,[[34]](#footnote-34) mit denen das Korpus der Umfeldbriefe bereits in einem anderen Kontext untersucht wurde.[[35]](#footnote-35) Alternativ wird das Vorkommen von Themen oder Marken auf Basis von Erwähnungen (engl. mentions), Hashtags oder speziell definierten Keywords untersucht. Für die thematische Analyse des Umfeldkorpus wurden die manuell durch die Bearbeiterinnen der Editionen vergebenen Schlagworte ausgewertet. Berücksichtigt man in der Analyse nur die erste Ebene der Thementaxonomie, dann gibt es bis dato 59 verschiedene Schlagworte, die bis dato insgesamt 2867mal vergeben wurden, womit jedem Brief im Schnitt rund 2,5 Schlagworte zugeordnet sind. Die zehn am häufigsten vergebenen Themen sind *Reisen* (247), *Krankheit bzw. Gesundheitszustand* (199), *Verlage / Verlegerisches* (169), *Familie/n* (152), *Finanzen* (144), *Briefverkehr* (141), *Kinder* (137), *Festlichkeiten* (126), *Literarisches Leben und Schaffen* (118) und *Freundschaften* (114).

Möchte man die Zahl der Themen in Bezug zu Jahren setzen, um so das Auf- oder Abflammen bestimmter Kommunikationsinhalte zu untersuchen, genügt es nicht, in absoluten Zahlen bzw. Summen zu rechnen, denn die Zahl der Briefe und deren Textlänge variiert stark in den verschiedenen Jahren. In Zeiträumen, für die besonders viele Briefe überliefert und in die Editionen aufgenommen sind, wie beispielsweise die Jahre 1808–1811, sind folglich auch mehr Themen verzeichnet (Abb. 1). Man braucht also eine relative Metrik, wie den *Share of Voice,* der Auskunft über anteilige Sichtbarkeit und Relevanz einer Marke oder eines Thema in den sozialen Medien gibt.[[36]](#footnote-36) Dazu werden Berichtzeiträume festgelegt, in denen der prozentuelle Anteil der Erwähnung einer Marke in Bezug auf alle erwähnten Marken berechnet wird, d. h. die Ermittlung des Share of Voice erfolgt in relativen Zahlen, wodurch die Unausgewogenheit an Masse und Umfang der Briefe im Korpus ausgeglichen wird.

Um die Analyse übersichtlich zu gestalten, wurden exemplarisch die drei am häufigsten vertretenen Themen im Umfeldkorpus – *Reisen, Krankheit bzw. Gesundheitszustand* und *Verlage / Verlegerisches* – ausgewertet. Die drei Themen wurden für die Briefe aus den Jahren 1800­­­­­–1826 insgesamt 911-mal von den Bearbeitern als Schlagwort vergeben, wobei die Vergabe von Schlagworten in Briefen ohne Datum in der Analyse nicht berücksichtigt wurde.



**### Abb. 3: Share of Voice der drei am häufigsten vergebenen Themenschlagworte. ###**

Abbildung 3 veranschaulicht den Share of Voice für die genannten Themen. Bei genauerer Betrachtung fällt auf, dass das Thema *Verlage / Verlegerisches* nicht durchgängig im Korpus vertreten ist und darüber hinaus eine Intensivierung des Thema um 1825/26 vorliegt. Reisen ist kontinuierlich über die Jahre ein Thema, allerdings sind dabei ebenfalls An- und Abstiege des SoV zu beobachten. Im Jahre 1812 könnte die vergleichsweise geringe Thematisierung von Reisen an den dato stattfindenden Befreiungskriegen liegen, durch welche die Reisetätigkeiten der Gesellschaft grundsätzlich eingeschränkt wurden. Ebenfalls 1812 sind Krankheiten bzw. Gesundheitszustand ein stark ausgeprägtes Thema, was..

Man muss den Share of Voice nicht zwangsläufig auf das Gesamtkorpus beziehen, sondern könnte auch einzelne Korrespondenzen in den Blick nehmen oder die thematischer Ausrichtung verschiedener Verfasserinnen vergleichen. Auch könnte man andere Berichtszeiträume anlegen.

Opinion

Sentimentanalyse ist ein Teilgebiet des Natural Language Processing, welches sich mit der Untersuchung der Intensität der in einem Text ausgedrückten Gefühle befasst. Die Methode eignet sich, um die Gefühle bzw. Tonalität von Usern gegenüber einem Thema oder einer Marke zu untersuchen.[[37]](#footnote-37) Die Datengrundlage für Sentiment Analysis sind für gewöhnlich „Rohdaten“, die wiederum durch Metadaten kontextualisierbar werden können. Auf Basis des Textdatesets,

Analysiert wurden Brieftexte, die mit Themen verschlagwortet wurden, die im Gesamtkorpus mindestens für mehr als 50 Briefe vergeben wurden. Dies trifft für 27 der 59 Themen zu.

Fazit

Es handelt sich bei den ersten beiden Analysen um relativ simple Berechnungen, die aber unser Verständnis über das Korpus erweitern können. Die Daten geben alles her, man muss nur damit arbeiten.

Doch die entscheidende Frage ist: welche Bücher befinden sich in dem Archiv und welche nicht? Die Häufung eines bestimmten Begriffs zu einer bestimmten Zeit kann immer noch an wenigen Autoren oder Einzelwerken hängen. Für auf den ersten Blick überraschende Treffer, finden bei der Überprüfung der einzelnen Fundstellen einfache Erklärungen. Doch trotz solcher „Kinderkrankheiten“ und blinder Flecken ist das enorme Potenzial, das in Big-Data-Analysen für die Geistes- und Kulturwissenschaften offensichtlich.

Gleichzeitig gilt es zu diskutieren, ob die ‚Social-Media-haftigkeit‘ der Briefe aus dem Umfeld Jean Pauls ausreicht, um mittels quantitativer Verfahren zu aussagekräftigen Ergebnissen zu gelangen.

Learnings:

Unschärfe

Masse an Daten vs. Verstehbarkeit der Visualisierung

Grundsätzlich zielen die Überlegungen des Beitrags nicht auf eine bloße Quantifizierung der Kommunikation, sondern vor allem auf eine Diskussion der Frage, wie man mit den qualitativ erschlossenen Editionsdaten zu wissenschaftlich fundierten quantitativen Analysen gelangen kann, und ob und wie diese die Forschung unterstützen und neue Perspektiven auf das Korpus generieren können. Gleichzeitig gilt es zu diskutieren, ob die ‚Social-Media-haftigkeit‘ der Briefe aus dem Umfeld Jean Pauls ausreicht, um mittels quantitativer Verfahren zu aussagekräftigen Ergebnissen zu gelangen.

Wir können von den Sozialen Medien nicht nur Methoden borgen, sondern sie auch für die Wissenschaftskommunikation nutzen.

Sugimoto et al. 2017, S. 2037

Sugimoto, Cassidy R., Sam Work, Vincent Larivière und Stefanie Haustein (2017): Scholarlyuse of social media and altmetrics: a review of the literature, in:Journal of the Associationfor Information Science and Technology, Jg. 68, Nr. 9, S. 2037-2062. https://doi.org/10.1002/asi.23833.

1. Zur Debatte siehe u. a. danah boyd, Kate Crawford, Critical Questions for Big Data, in: Information, Communication & Society 15:5 (2012), S. 662–679, hier S. 663; Stieglitz S, Brockmann T, Dang-Xuan L (2012). Usage of social media for political communication. In: Proc of 16th Pacific Asia conference on information systems, Ho Chi Minh City, Vietnam; Zeng D, Chen H, Lusch R, Li S (2010) Social media analytics and intelligence. IEEE Intelligent Systems 25(6):13–16. [↑](#footnote-ref-1)
2. Stieglitz, S., Dang-Xuan, L. Social media and political communication: a social media analytics framework. *Soc. Netw. Anal. Min.* 3,1277–1291 (2013), hier S. 90. [↑](#footnote-ref-2)
3. Moretti, Franco (2000). "Conjectures on World Literature". New Left Review. 1. Jänicke, Stefan, Franzini Greta, Cheema Muhammad Faisal, and Gerik Scheuermann: On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges, Eurographics Conference on Visualization (EuroVis) - STARs 2015: <https://www.informatik.uni-leipzig.de/~stjaenicke/Survey.pdf> [↑](#footnote-ref-3)
4. Quantitative und statistische Analysen kommen bisher im Kontext digitaler Editionen eher selten zum Einsatz. ###Schleiermacher, Selbstzitation### [↑](#footnote-ref-4)
5. Für zwei weitere Beschäftigungen mit Jean Pauls Umfeld siehe die Beiträge von Selma Jahnke und Michael Rölcke in diesem Band. [↑](#footnote-ref-5)
6. Briefe aus Jean Pauls Umfeld. Herausgegeben an der Berlin-Brandenburgischen Akademie der Wissenschaften von Selma Jahnke und Michael Rölcke (2020–2022), in: Jean Paul - Sämtliche Briefe digital. Herausgegeben im Auftrag der Berlin-Brandenburgischen Akademie der Wissenschaften von Markus Bernauer, Norbert Miller und Frederike Neuber (2018–2020). URL: [http://jeanpaul-edition.de](http://jeanpaul-edition.de/brief.html?num=VI_3). [↑](#footnote-ref-6)
7. Die genaue Definition der Begriffe erfolgt im Kapitel zur jeweiligen Analyse. [↑](#footnote-ref-7)
8. Für die Datensätze der Version 5 der digitalen Edition siehe telota/jean\_paul\_briefe: Daten der Briefe von Jean Paul und der Briefe aus seinem Umfeld (v.5.0). Zenodo. https://doi.org/10.5281/zenodo.6322839 [↑](#footnote-ref-8)
9. Markus Strohmaier und Maria Zens: Analyse Sozialer Medien an der Schnittstelle zwischen Informatik und Sozialwissenschaften, in: Soziale Medien: Gegenstand und Instrument der Forschung, Schriftenreihe der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute (2014), S. 73-95, hier S. 73. [↑](#footnote-ref-9)
10. Freilich gibt es weitere Parallelen zwischen digitalen und epistolaren Netzwerken, die jedoch im Kontext dieses Beitrags nur marginal relevant sind, darunter multipolare Korrespondenzstrukturen, d. h. Nachrichten von und an mehrere Personen, sowie die Entgrenzung von Privatheit und Öffentlichkeit, die aber in vielen anderen Beiträgen des Bandes thematisiert werden. [↑](#footnote-ref-10)
11. Zu Editionen als Daten siehe u.a. Patrick Sahle: Zwischen Mediengebundenheit und Transmedialisierung. Anmerkungen zum Verhältnis von Edition und Medien. In: editio, Bd. 24. Berlin 2010, S. 23–36. [↑](#footnote-ref-11)
12. Selbstzitation; Anreicherung!!! Frederike Neuber: Der Text und seine (Re)Produktion, De Gruyter, 2023. [↑](#footnote-ref-12)
13. Strohmaier und Zens 2014, hier S. 74. [↑](#footnote-ref-13)
14. Schöch, Christof. “Big? Smart? Clean? Messy? Data in the Humanities.” *Journal of the Digital Humanities* 2, no. 3 (2013): 2–13, hier S. 4. URL: <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>. Ähnlich beschreiben es boyd und Crawford (2012) in ihrer Gegenüberstellung von ‘big data’ und ‘small data’, S. 670: “[I]t is increasingly important to recognize the value of ‘small data’. Research insights can be found at any level, including at very modest scales.” [↑](#footnote-ref-14)
15. Word Wide Web Consortium w3c: Extensible Markup Language 1.0, URL: <<https://www.w3.org/TR/xml/>>; Text Encoding Initiative: P5: Guidelines for Electronic Text Encoding and Interchange (Version 4.3.0). URL: <<http://www.tei-c.org>>. [↑](#footnote-ref-15)
16. DTABf. Deutsches Textarchiv – Basisformat (2011–2020), hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. <http://deutschestextarchiv.de/doku/basisformat>. [↑](#footnote-ref-16)
17. ediarum/ediarum.BASE.edit, Version 2.0, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften, bearbeitet von Stefan Dumont, Nadine Arndt, Sascha Grabsch und Lou Klappenbach. 2011-2022. URL: <<https://github.com/ediarum/ediarum.BASE.edit>> [↑](#footnote-ref-17)
18. Verweis Schöch… [↑](#footnote-ref-18)
19. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd\_node.html (zuletzt aufgerufen am 24.02.2022). [↑](#footnote-ref-19)
20. https://www.geonames.org/ (zuletzt aufgerufen am 24.02.2022). [↑](#footnote-ref-20)
21. https://www.iso.org/home.html (zuletzt aufgerufen am 24.02.2022). [↑](#footnote-ref-21)
22. Stefan Dumont, Sascha Grabsch und Jonas Müller-Laackman (Hg.): correspSearch – Briefeditionen vernetzen (2.0.0) [Webservice]. Berlin-Brandenburgische Akademie der Wissenschaften 2021. https://correspSearch.net [↑](#footnote-ref-22)
23. Beide Kategoriesysteme, Korrespondenzkreise und Themen, fungieren in der digitalen Edition als Register, über die man auf die Briefe zugreifen kann und die Gemeinsamkeiten zwischen Briefen, auch wenn sie unterschiedlicher Teilkorrespondenzen entstammen, sichtbar machen. Die Verschlagwortung entsteht parallel zur Erschließung der Edition, d.h. sie kann sich bei wachsender Briefmasse verändern oder erweitern. [↑](#footnote-ref-23)
24. boyd und Crawford 2012, S. 670. [↑](#footnote-ref-24)
25. Stieglitz, S., Dang-Xuan, (2013). URL: https://doi.org/10.1007/s13278-012-0079-3, u.a. Fig. 4. Die Autoren entwickelten das Framework für Social Media Analytics im Kontext von politischer Kommunikation, wobei das Modell durch das konzeptionelle Abstraktionslevel generalisierbar für alle Formen der Datenanalyse gelten kann. [↑](#footnote-ref-25)
26. ####Ähnliche Beiträge aus den DH#### [↑](#footnote-ref-26)
27. Dies ist bei der Analyse der Umfeldbriefe sowie in allen anderen Anwendungsfällen, in denen komplexe Sachverhalte in Daten analysiert werden, nicht zu erwarten. Andere Grafik!!!! [↑](#footnote-ref-27)
28. Link zu den Ergebnissen. [↑](#footnote-ref-28)
29. Siehe <https://www.deutschestextarchiv.de/demo/cab/> und Jurish, B. *Finite-state Canonicalization Techniques for Historical German.* PhD thesis, Universität Potsdam, 2012. URN [urn:nbn:de:kobv:517-opus-55789](https://nbn-resolving.org/resolver?identifier=urn%3Anbn%3Ade%3Akobv%3A517-opus-55789). [↑](#footnote-ref-29)
30. Schmidt, T., Dangel, J. & Wolff, C. (2021). SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities. In: Schmidt, T. & Wolff, C. (Eds.), Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021). Glückstadt: Verlag Werner Hülsbusch. (pp. 156—172). DOI: 10.5283/epub.44943 [↑](#footnote-ref-30)
31. ZOTERO - Marketing: Eine managementorientierte Einführung, hier S. 310. [↑](#footnote-ref-31)
32. Gleiches gilt übrigens für die Empfangskontakte (direkter Empfang und Mitleserschaft zusammengezählt), die mit 1435 noch deutlich höher als die Sendekontakte ausfallen.­­­ [↑](#footnote-ref-32)
33. Anne Marx, Media für Manager: Was Sie über Medien und Media-Agenturen wissen müssen, 2012, hier S. 95–96. [↑](#footnote-ref-33)
34. Vgl. u.a.Dimitrios Milioris: Topic Detection and Classification in Social Networks: The Twitter Case. Springer International Publishing 2017, S. 13. [↑](#footnote-ref-34)
35. Ulrike Henny-Krahmer und Frederike Neuber, Topic Modeling in Digital Scholarly Editions, in: Geiger, Bernhard, Ulrike Henny-Krahmer, Fabian Kaßner, Marc Lemke, Gerlinde Schneider und Martina Scholger (Hrsg.): Machine Learning and Data Mining for Digital Scholarly Editions. Schriften des Instituts für Dokumentologie und Editorik 18. Norderstedt: Books on Demand. Forthcoming. [↑](#footnote-ref-35)
36. Christoph Burmann, Tilo Halaszovich, Michael Schade, Rico Piehler: Identitätsbasierte Markenführung. Springer 2018. S. 266 [↑](#footnote-ref-36)
37. Carlos A. Iglesias und Antonio Moreno: Editorial, in: dies.: Sentiment Analysis for Social Media. 2020. S. 1-4. [↑](#footnote-ref-37)