Die Briefe aus dem Umfeld Jean Pauls als Social Media-Corpus

# Einleitung

Wenn die Briefkultur um 1800 Ähnlichkeiten zu den Social Media von heute aufweist, dann sollten die Strukturen und Praktiken der epistolaren Kommunikation 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 Korpora der Social Media vorrangig aus quantitativer Perspektive analysiert, u. a. von Sozialwissenschaften, Kommunikationswissenschaften und (Wirtschafts-)Informatik.[[1]](#footnote-1) Ein solches ‘distant reading’ von historischen Korrespondenzen, das sich von der Detaillektüre einzelner Texte und deren Inhalten löst, kann eine abstrakte bzw. übergeordnete Perspektive auf einen größeren Briefbestand generieren, wodurch allgemeine Merkmale und Muster der Kommunikation identifizierbar werden.[[2]](#footnote-2) Ausgehend von diesen Überlegungen und am Beispiel einer Social Media Community um 1800, dem Korrespondenzumfeld des Schriftstellers Jean Paul, stellt der Beitrag zunächst Parallelen zwischen digitalen und epistolaren Communities heraus. Im Folgenden wird das Potential der Editionsdaten für quantitative Analysen diskutiert, bevor Methode und Workflow der Analyse vorgestellt werden. Die eigentlichen Analysen, deren Ergebnisse erläutert, visualisiert und diskutiert werden, nehmen vorrangig drei Untersuchungsfelder in den Blick: Performance und Akteur:innen im Korpus, Themen und Trends in den Brieftexten sowie Tonalität und Stimmung der Kommunikation. Grundsätzlich zielen die Überlegungen des Beitrags nicht auf eine bloße Quantifizierung der Kommunikation, sondern vor allem auf die 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 bieten 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.

# 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. Beide Netzwerke sind ‘Plattformen’, auf denen sich *Communities*, d. h. Gruppen von Individuen mit gemeinsamen Interessen, zusammenfinden. Ein epistolare Community um 1800 ist das Umfeld des Schriftstellers Jean Paul (1763–1825), bestehend aus Familie, Freunden und Kollegen. Seit 2020 erscheint ein Korpus der Umfeldbriefe als Teil der Edition *Jean Paul - Sämtliche Briefe digital,* das zum jetzigen Zeitpunkt 1156 Dokumente umfasst (Stand Mai 2022, v.5.0).[[3]](#footnote-3)

Zur Edition der Umfeldbriefe zählen derzeit 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, dem Jurist und Schriftsteller Christian Otto und dem vielvernetzten jüdischen Kaufmann Emanuel (Osmund). Abgesehen von dieser personenbezogenen Sicht auf das Korpus, lassen sich verschiedene Kommunikationszirkel ausmachen, wie die Kommunikation der Richters mit Jean Pauls und Carolines Kindern Emma (1802–1853), Max (1803–1821) und Odilie (1804–1865) oder das Korrespondenznetz, dass sich während Jean Pauls Zeit in Weimar bildete, dem u.a. Caroline und Johann Gottfried Herder angehören. Der Kitt dieser Communities sind gemeinsame Bekanntschaften, Interessen und Themen. Wie User auf Instagram unter den Hashtags *#travelgram* oder *#instatravel* von ihren Reisen berichten, ergeben sich auch im Umfeld Jean Pauls thematische Bündelungen der Briefe, beispielsweise 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 gestaltet die Community bzw. die jeweiligen *User* die Inhalte bzw. den *Content* 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.”[[4]](#footnote-4) Zwar legen die großen Briefeditionen mit ihrem Fokus auf ‘geschichtsträchtige’ Individuen nahe, dass in historischen Netzwerken nur ‘intellektueller’ Content verhandelt wurde, allerdings - und auch das zeigt die Edition der Umfeldbriefe Jean Pauls - spannten sich Netzwerke um die schreibenden Protagonisten weiter in deren Orbit, in dem sich eigenständige Communities mit eigenen Themen bildeten. Im Zentrum von Netzwerken gibt es damals wie heute meist ein oder mehrere personelle ‘Epizentren’, die besonders aktiv an der Kommunikation teilhaben und Interaktionen bündeln, und die man heutzutage als *Influencer* bezeichnet. Eine zentrale Figur in Jean Pauls Umfeld bzw. im edierten Korpus ist u.a. seine Frau Caroline Richter.[[5]](#footnote-5)

Die Umfeldbriefe weisen nicht nur Parallelen zu Inhalten und Praktiken der digitalen Kommunikation von heute auf, man kann auch ähnliche Fragen an sie richten, darunter “Wer partizipiert im Netzwerk und welche Communities bilden sich?”, “Welcher Content liegt generell im *Trend* oder geht zu bestimmten Zeitpunkten gar viral?” und “Wie wird über bestimmte Dinge gesprochen?”

# Digitale Edition als *Smart Data*

Sowohl bei Social Media-Inhalten als auch bei digitalen Editionstexten handelt es sich im Kern um Datensätze,[[6]](#footnote-6) die sich allerdings hinsichtlich einiger Aspekte grundlegend unterscheiden. Datensätze sozialer Netzwerke kann man, insofern sie frei verfügbar sind, über technische Schnittstellen oder über spezielle Monitoring-Tools beziehen, wobei, letztere vor allem Dingen von Unternehmen zur Auswertung von Interaktionen oder Inhalte verwendet werden, um zukünftige Entwicklungen wie Trends und Kundeninteressen zu prognostizieren. Auch wenn die Daten der Social Media für die Auswertung vorverarbeitet werden, indem beispielsweise Metadaten extrahiert werden, ist die Überlieferung selbst der Analysedatensatz. Im Fall von historischen Korrespondenzen wird die Überlieferung hingegen durch die digitale Edition bzw. die Kodierung repräsentiert, wobei editorische Entscheidungen über die Erfassung und Erschließung des Textes zum Tragen kommen. Beispielsweise können sich Editorinnen entscheiden, bestimmte Textbestandteile gar nicht erst zu erfassen oder bestimmte textkritische Phänomene zu normalisieren. Gleichzeitig ist die editorische Bearbeitung aber auch immer eine Wissensanreicherung, durch welche die Überlieferung bzw. die Daten an Informationen gewinnen, beispielsweise durch die Referenzierung von Personennamen auf Normdatensätze.

Daten aus Social Media-Plattformen und 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,”[[7]](#footnote-7) 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.[[8]](#footnote-8)

Mit 1156 Dokumenten ist das Korpus der Umfeldbriefe sowohl aus Sicht der Social Media Analytics als auch aus Perspektive der quantitativen Textanalyse ein verhältnismäßig kleines, aber ein “intelligentes” Datenset. Die Briefe liegen im W3C-Standardformat XML vor und sind nach den Richtlinien der Text Encoding Initiative (TEI), dem de facto Standard zur Kodierung wissenschaftlicher digitaler Editionen, erfasst.[[9]](#footnote-9) Das Kodierungsschema der Umfeldbriefe orientiert sich an zwei TEI-Subsets: dem Basisformat des Deutschen Textarchivs[[10]](#footnote-10) und ediarum.BASE, das der Software ediarum,[[11]](#footnote-11) 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.[[12]](#footnote-12)

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[[13]](#footnote-13) verknüpft, Orte mit Identifikatoren von GeoNames.[[14]](#footnote-14) Datumsangaben sind einheitlich im Format der *International Organization for Standardization* (ISO)[[15]](#footnote-15) 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,[[16]](#footnote-16) 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 Schlagwörtern 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.[[17]](#footnote-17) 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 Schlagwörtern 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 auch allgemein für Sozial- und Kulturwissenschaften oder auch 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.”[[18]](#footnote-18)

# Methode und Workflow

Um von Daten zu Analyseergebnissen zu kommen, sind verschiedene Schritte nötig: In ihrem „Social Media Analytics Framework“[[19]](#footnote-19) definieren Stieglitz und Dang-Xuan ein fünfstufiges Verfahren der Datenanalyse von Social-Media-Korpora, das sich wie folgt zusammensetzt: 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.

Das Modell von Stieglitz und Dang-Xuan entspricht in den Grundzügen dem Vorgehen bei der Analyse der Umfeldbriefe, allerdings mit zwei Einschränkungen. Zum einen beschränken sich die quantitativen Analysemethoden durch die Masse und den geringen Strukturierungsgrad von Social Media-Daten vorrangig auf Verfahren des maschinellen Lernens, während im Falle der Analyse der stark strukturierten Informationen im Datensatz der Umfeldbriefe auch konkrete Datenfelder (z.B. Schlagwörter) statistisch auswertbar sind. Zum anderen legt das Modell nahe, der Workflow verläuft linear von der Eingabe eines Datensatzes bis zur Ausgabe der Ergebnisse. Dies ist bei der Analyse der Umfeldbriefe sowie in allen anderen Anwendungsfällen, in denen komplexe Sachverhalte in Daten analysiert werden, nicht zu erwarten. Vielmehr muss man sich die ‘Laufrichtung’ zwischen Preprocessing (iii.), Analyse (iv.) und Ergebnis (v.) als Spirale vorstellen, bei der sich mit jeder Drehung das Verständnis über die Analyse selbst erweitert: 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 nun 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 überladen und nicht ‘lesbar’ ist. In solch einem Fall würde man Analysedaten oder Methode ebenfalls entsprechend anpassen, um den Fokus der Analyse stringenter herauszustellen.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

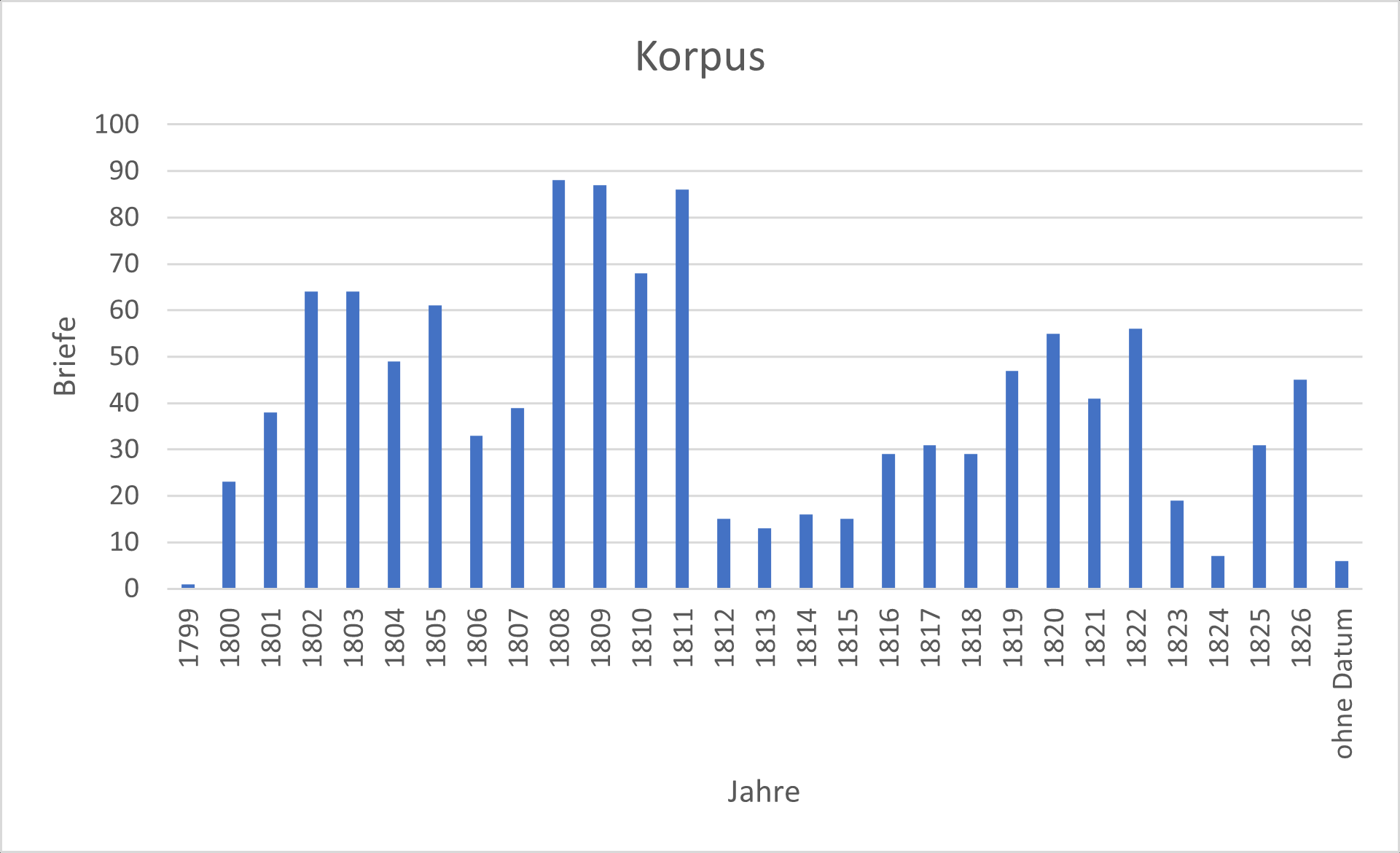
Abbildung 1 Workflow der Datenanalyse.

Abbildung 1 zeigt den Workflow bei der Analyse der Umfeldbriefe. Die drei Fragenbereiche nach dem „Wer?“, dem „Worüber?“ und dem „Wie?“ der Kommunikation bilden den Ausgangspunkt. Die vorliegenden Editionsdaten werden im Prozess des Preprocessing auf die Aspekte reduziert, die für die Analysen relevant sind. Erstellt wurden zwei Analysedatensets:

* Datenset I – im Folgenden „Metadatenset“ – besteht im Wesentlichen aus den editorischen Informationen, mit denen die Ursprungstexte angereichert wurden. Der XML-Datensatz umfasst briefweise Titeldaten, Korrespondenzmetadaten (Sender:in, Empfänger:in, Sendedatum, Sendeort), Schlagwörter zu Korrespondenzkreisen und Themen sowie Listen der Personen, Orte und Werke, über die im Brieftext indiziert sind, d.h. über die gesprochen wurde. Für die nachfolgenden Analysen wurde das Metadatenset mit XSLT prozessiert und Ergebnisse im Tabellenformat csv generiert, welche dann mit verschiedenen Tools für Visualisierungen aufbereitet wurde.[[20]](#footnote-20)
* Datenset II – im Folgenden „Textdatenset“ – besteht aus dem Brieftext, dessen historische Orthografie und Varianz mit Hilfe der Software CAB[[21]](#footnote-21) normalisiert wurde, um textbasierte Analysen zu vereinfachen. Editorische Anreicherungen wurden gänzlich entfernt. Das Datenset wurde als Grundlage der Analyse nach der Tonalität der Briefkommunikation verwendet, die mit dem Tool SentText im Verfahren der Sentiment Analysis ermittelt wurde. [[22]](#footnote-22)

# Analysen

Bevor im Anschluss die eigentlichen Analysen vorgestellt werden, soll noch auf einen Sachverhalt hingewiesen werden, der für die Deutung der Ergebnisse relevant ist: Bei den edierten Umfeldbriefe handelt es sich um kein abgeschlossenes Korpus, und zwar unter zweierlei Gesichtspunkten: 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 Editor:innen 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 Editor:innen, 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 und nicht aber auf die Wirklichkeit zu beziehen. Wenn man an das Korpus also beispielsweise die Frage richtet, wer im Umfeld wie viel geschrieben hat, dann fragt man genau genommen nach dem Anteil der Briefe eines einzelnen Korrespondenten am derzeit verfügbaren Korpus, d.h. an den Briefen, die ausgewählt wurden und überliefert sind.

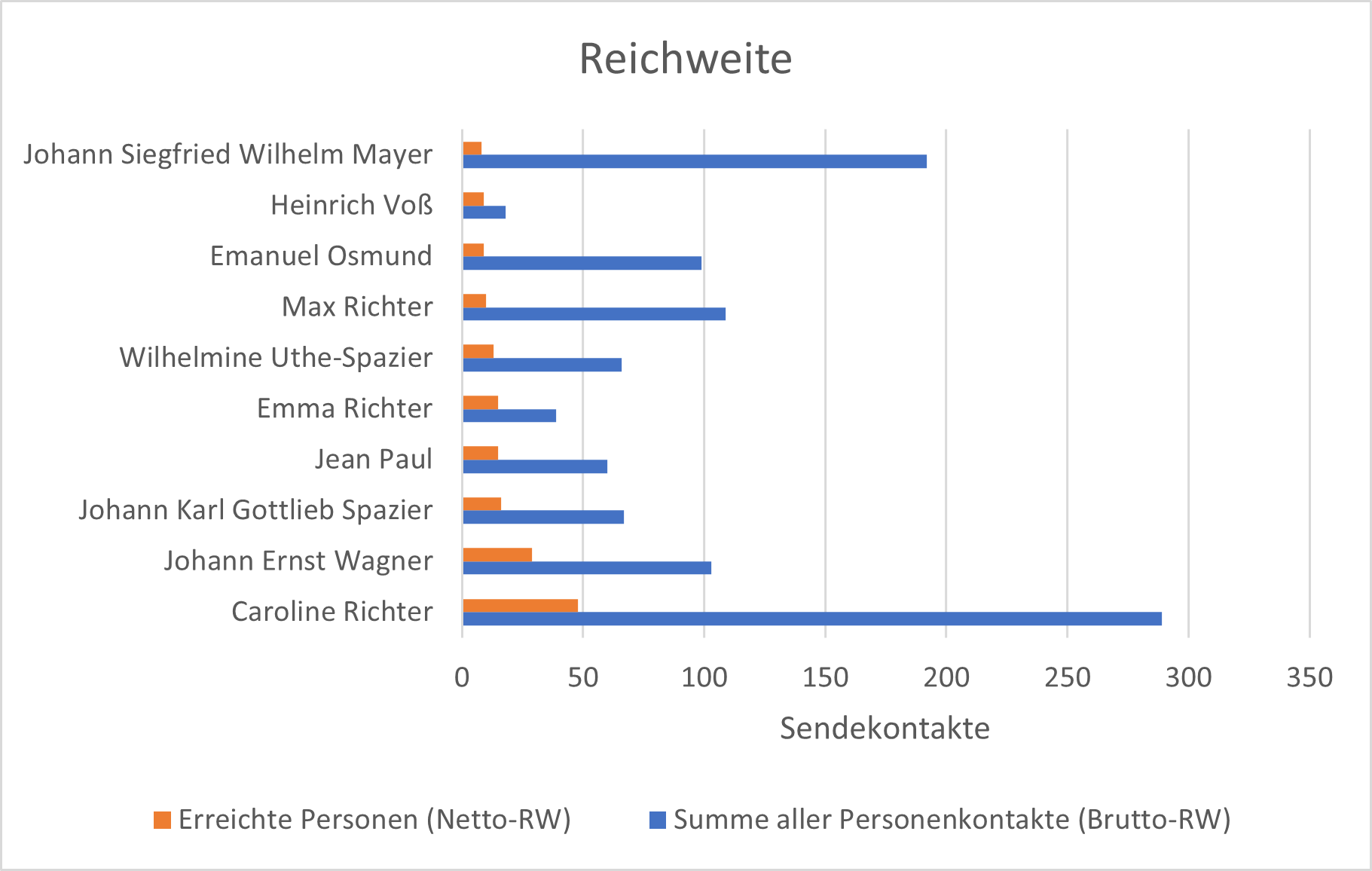


## Rollen und Reichweiten

Bei der Erhebung der Metadaten wurden verschiedene Rollen bestimmt, mit denen Personen an der Briefkommunikation beteiligt sein können: Verfasser:innen, im Folgenden Sender:innen, Empfänger:innen und Personen, die Briefe lesen, auch wenn diese nicht explizit für sie bestimmt waren. Ein Spezifikum der Umfeldbriefe ist, dass Briefe kollaborativ verfasst sein können und resp. oder nicht nur eine, sondern mehrere Empfänger haben können. Diese multipolaren Kommunikationsstrukturen sollen bei den nachfolgenden Berechnungen berücksichtigt werden, weshalb jedes Verfassen eines Briefes – ob allein oder gemeinschaftlich – als eigener Sendekontakt gewertet wird, mit dem die 109 verschiedenen Verfasser:innen sich in das Kommunikationsnetzwerk einbringen. Bedingt durch das kollaborative Verfassen von Briefen liegt die Summe aller Sendekontakte mit 1247 höher als die Summe der Briefe (1156). Gleiches gilt für die Empfgangskontakte der 129 Empfänger:innen und Mitleser:innen, die für die Zählung zusammengelegt wurden, die mit 1435 noch deutlich höher als die Sendekontakte ausfallen, wobei allein 259 dieser Empfangskontakte die erweiterte Leserschaft abbilden.

Betrachtet man die Sender:innen-Seite, so hat im Schnitt jede Korrespondent:in rund 11 mal einen eigenständigen oder einem kollaborativen Brief verfasst. Sieht man genauer hin, dann haben von den 109 Verfasser:innen aber lediglich 23 Personen 11 oder mehr Briefe verfasst, d.h. rund 21%. Allein die 5 aktivsten Verfasser:innen, machen mit 792 Kontaktaufnahmen 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 Kontaktaufnahmen bzw. Briefe angeht, im derzeitigen Datenbestand eine enorme Ballung allein fünf Verfasser:innen 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 veranschaulichen und zu verstehen, aber sie sagt nur in Ansätzen etwas über die Rolle der einzelnen Sender:innen im Gesamtkorpus aus. Es bietet sich daher an, in einer weiteren Analyse zwischen der Summe aller Sendekontakte einer Person, bei der Überschneidung von gleichen Empfängern nicht berücksichtigt werden, und der Zahl der der verschiedenen Empfänger:innen, die mit dem Sendekontakt erreicht wurden, zu unterscheiden. Ähnlich funktioniert die Berechnung der Reichweite einer Website oder Marke im Social Media Marketing, bei der die Zugriffe auf eine Webseite nach der Summe aller Zugriffe und der Summe der Erstzugriffe differenziert werden, um zwei Kennzahlen zu ermitteln: Brutto- und Nettoreichweite. Abbildung 1 zeigt die 10 Sender:innen mit der höchsten Nettoreichweite, also mit den meisten Sendekontakten mit verschiedenen Empfängern, und stellt diesen Wert in Bezug zur Bruttoreichweite, d.h. zu allen Sendekontakten, die von dieser Person ausgehen.



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?

## Themen und Trends

Bei der Ermittlung und Auswertung von Themen werden im Bereich der Social Media Analysen meist Verfahren des maschinellen Lernens wie Topic Modeling eingesetzt.[[23]](#footnote-23) Da die Umfeldbriefe im Rahmen der Edition von den Bearbeiter:innen nach bestimmten Themen klassifiziert werden, wird für nachfolgende Analysen diese qualitative Verschlagwortung herangezogen. Die 59 Themen der manuellen Themenklassifikation sind bis dato 2867 briefweise vergeben worden, womit jedem Brief im Schnitt rund 2,5 Schlagwörter 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. Um die Rolle eines Themas in der Umfeldcommunity zu messen und im zeitlichen Verlauf zu betrachten wird die Metrik „Share of Voice“ herangezogen. Es handelt sich dabei um eine Kennzahl, die Auskunft über Sichtbarkeit und Relevanz einer Marke oder eines Thema in den sozialen Medien gibt. Dazu werden Berichtzeiträume festgelegt, in denen der prozentuelle Anteil eines Themas in Bezug auf alle erwähnten Themen 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 Auswertung überschaubar zu gestalten, wurden für die Analyse exemplarisch die drei am häufigsten vertretenen Themen im Umfeldkorpus – *Reisen*, *Krankheit bzw. Gesundheitszustand* und *Verlage / Verlegerisches* – ausgewählt. Sie wurden für die Briefe aus den Jahren 1800-1826 insgesamt 911 mal von den Bearbeitern als Schlagwort vergeben, wobei die Vergabe von Schlagwörtern in Briefen ohne Datum in der Analyse nicht berücksichtigt ist.

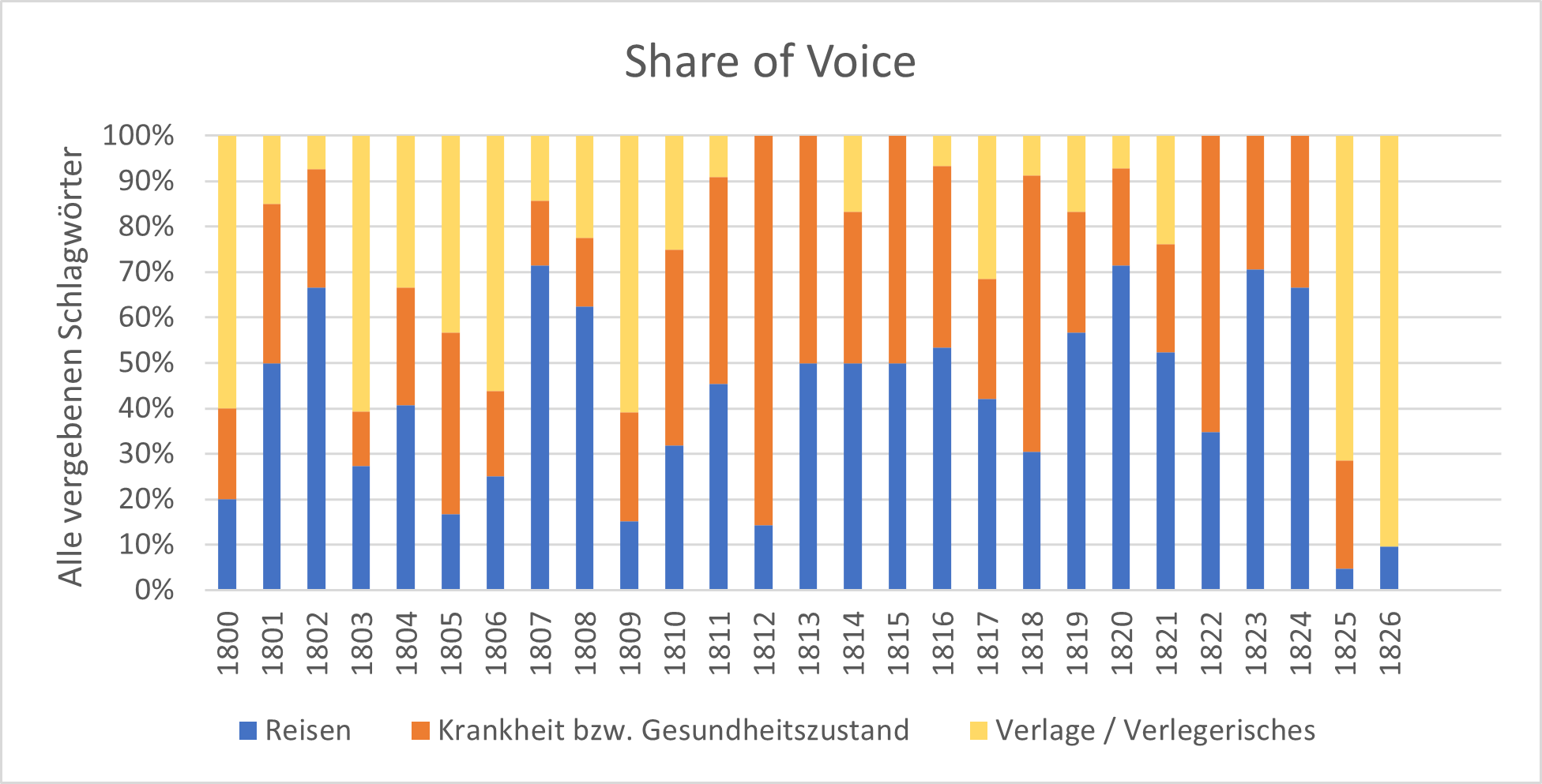


Abbildung 2 veranschaulicht den Share of Voice für die genannten Themen. Dabei fällt unmittelbar ins Auge, dass das Thema Verlage nicht durchgängig im Korpus vertreten ist und darüber hinaus eine Intensivierung des Thema um 1825/26 vorliegt.

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 grundsätzlich eingeschränkt wurden. Ebenfalls 1812 sind Krankheiten bzw. Gesundheitszustand ein stark ausgeprägtes Thema, was..

## Tonalität und Stimmung

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.[[24]](#footnote-24) 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

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

1. danah boyd & Kate Crawford: Critical Questions for Big Data, in: Information, Communication & Society 15:5 (2012), S. 662-679, S. 663. DOI: 10.1080/1369118X.2012.678878. Gleiches gilt für das Methodenset der *Social Media Analytics*, mit dem u.a. politische Akteure und Unternehmen kontinuierlich Informationen aus sozialen Medien sammeln, überwachen, analysieren und visualisieren, um das Verhalten potentieller Wähler oder Kunden zu prognostizieren und zu steuern. Stieglitz, S., Dang-Xuan, L. Social media and political communication: a social media analytics framework. *Soc. Netw. Anal. Min.* **3,** 1277–1291 (2013). https://doi.org/10.1007/s13278-012-0079-3 S. 1277. [↑](#footnote-ref-1)
2. 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> Die digitale Literaturwissenschaft arbeiten unter dem von Franco Moretti eingeführten Schlagwort „distant reading“ bereits seit geraumer Zeit mit quantitativen Verfahren der Textanalyse, deren Anwendung im Kontext digitaler Editionen aber eher selten ist. Vgl. Franco Moretti, der ihn im Jahre 2000 in seinem Essay „Conjectures on World Literatur“ eingeführt hat. (Vgl. Moretti 2000). [↑](#footnote-ref-2)
3. Wenn man die ebenfalls digital publizierten Briefe von Jean Paul hinzu rechnet dann ist auf der Editionsplattform mit derzeit 6682 Briefen eines der umfangreichsten edierten Korrespondenzkorpora der Zeit um 1800 digital verfügbar. Siehe 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). Für die Datensätze der 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-3)
4. 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-4)
5. 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. [↑](#footnote-ref-5)
6. 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-6)
7. Strohmaier und Zens 2014, hier S. 74. [↑](#footnote-ref-7)
8. 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-8)
9. 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-9)
10. DTABf. Deutsches Textarchiv – Basisformat (2011–2020), hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. <http://deutschestextarchiv.de/doku/basisformat>. [↑](#footnote-ref-10)
11. 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-11)
12. Distant reading [↑](#footnote-ref-12)
13. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd\_node.html (zuletzt aufgerufen am 24.02.2022). [↑](#footnote-ref-13)
14. https://www.geonames.org/ (zuletzt aufgerufen am 24.02.2022). [↑](#footnote-ref-14)
15. https://www.iso.org/home.html (zuletzt aufgerufen am 24.02.2022). [↑](#footnote-ref-15)
16. 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-16)
17. 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-17)
18. boyd und Crawford 2012, S. 670. [↑](#footnote-ref-18)
19. Stieglitz und Dang-Xuan 2013, 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-19)
20. Für die Visualisierungen wurde sowohl die Diagrammfunktionen von Microsoft Excel als auch die Online-Tool raw.graphs verwendet. [↑](#footnote-ref-20)
21. 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-21)
22. 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-22)
23. Vgl. u.a.Dimitrios Milioris: Topic Detection and Classification in Social Networks: The Twitter Case. Springer International Publishing 2017, S. 13. [↑](#footnote-ref-23)
24. Carlos A. Iglesias und Antonio Moreno: Editorial, in: dies.: Sentiment Analysis for Social Media. 2020. S. 1-4. [↑](#footnote-ref-24)