Quantitative Analyse von Zeitungsartikeln und Online-Nachrichten

108

Harald Klein

Während Befragungen (Helfferich und Reinecke, Kapitel 55 und 62 in diesem Band) in den Sozialwissenschaften in jedem Methodenbuch umfangreich abgehandelt werden, gilt dies nicht für die Verfahren, mit denen Texte analysiert werden. In diesem Beitrag geht es darum, Artikel aus Online- und Printmedien (d. h. Zeitungen und Zeitschriften) als Datenmaterial für die quantitative Sozialforschung zu erschließen (zur qualitativen Analyse dieser Daten, Taddicken, Kapitel 107 in diesem Band).

108.1 Beschaffung

Die Beschaffung von Zeitungs- oder Zeitschriftenartikeln ist für aktuelle Ausgaben leicht, weil sie im Handel erhältlich sind. Alle anderen Ausgaben von Zeitungen oder Zeitschriften müssen entweder beim Verlag erworben oder in Archiven gesucht werden. Hierbei kann es sich um Verlagsarchive oder auch um frei zugängliche Archive handeln. Das Zeitungsarchiv der Westfälischen-Wilhelms Universität in Münster ist eines der größten Pressearchive im deutschsprachigen Raum, das für jedermann zugänglich ist. Aber auch an anderen Universitäten gibt es Pressearchive. Einen aktuellen Überblick gibt die Website http://www.textanalysis.info.

Oft sind Zeitungen und Zeitschriften nicht mehr in gedruckter Form zugänglich, weil sie zum einen sehr viel Platz benötigen und zum anderen gerade bei Zeitungen die Qualität des Druckbildes im Laufe der Zeit immer mehr abnimmt. Der Grund dafür ist qualitativ minderwertiges Papier, auf dem Zeitungen gedruckt werden und welches aus Kostengründen zunehmend verwendet wird – es muss auch nicht lange haltbar sein. Daher werden Zeitungen und Zeitschriften verfilmt, entweder auf Mikrofilm oder Mikrofiche.

N. Baur und J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung*, https://doi.org/10.1007/978-3-658-37985-8_108

108.2 Textaufbereitung

Bei textanalytischen Verfahren wird das Material codiert (Mayring/Fenzl, Kapitel 43 in diesem Band), das kann man konventionell mit Papierausdrucken machen, aber auch mit entsprechender Software, entsprechende Übersichten sind im Internet zu finden (Kuckartz/Rädiker, Kapitel 32 in diesem Band). Dazu muss das Material gegebenenfalls digitalisiert werden, dafür gibt es einige Möglichkeiten wie Abschreiben, Scannen, Diktieren. Werden Texte abgeschrieben, so ist dies mit hohem Zeit- und Kostenaufwand verbunden, so dass man Alternativen wie das Scannen oder Diktieren in Betracht ziehen sollte. Diktieren ist über längere Zeit gesehen anstrengend und für große Textmengen nicht geeignet, dafür aber sehr schnell. Die Anstrengung resultiert daraus, dass für eine gute Spracherkennung eine entsprechend gute Aussprache notwendig ist, denn diese bestimmt die Qualität der Umsetzung von Sprache in Text und ist nie ganz fehlerfrei. Der gesamte diktierte Text muss daher auf Fehler überprüft werden, so dass der Geschwindigkeitsvorteil der Spracherkennung gegenüber dem Abschreiben schrumpft. Gute Schreibkräfte schaffen mehr als 250 Zeichen pro Minute, sind meist schneller und machen weniger Tippfehler. Im Vergleich zur Spracherkennung muss auch nicht der gesamte Text Korrektur gelesen werden, weil Schreibkräfte wissen, ob und an welcher Stelle im Text sie Tippfehler gemacht haben.

Beim Scannen können insbesondere bei Zeitungen erhebliche Probleme auftreten, da die Druckqualität so schlecht ist, dass man mit Scannern und entsprechender Texterkennungssoftware erheblich langsamer ist als wenn der Text abgeschrieben würde (Klein 2013: 18–20). Bei Zeitschriften ist das anders, da das Papier erheblich besser und so der Einsatz von Scannern sinnvoll ist.

108.3 Datenauswahl

Bevor man anfängt die Artikel zu sammeln, sollte man sich darüber klar sein, ob man alle Artikel – also eine Vollerhebung – oder eine Stichprobe analysieren möchte. Diese wird weiterhin durch die zur Verfügung stehenden Ressourcen finanzieller bzw. personeller Art determiniert. Bevor man eine Stichprobe (Häder/Häder, Kapitel 27 in diesem Band) ziehen kann, muss die Grundgesamtheit definiert werden. Geht es bei einer Fragestellung beispielsweise darum, wie ein Thema bzw. Personen(gruppen) in deutschen Zeitungen dargestellt werden, so muss man wissen, dass nicht jede Zeitung eine eigene Redaktion hat – also als publizistische Einheit gezählt wird. Wird dies nicht berücksichtigt, besteht die Gefahr einer Gewichtung, wenn man beispielsweise mehrere Ausgaben der WAZ in die Stichprobe mit einbezieht. Bei der WAZ bietet es sich an, die Hauptausgabe Essen zu wählen.

Die publizistischen Einheiten in Deutschland verringerten sich im Verlauf der letzten Jahrzehnte und erreichten 2018 eine Anzahl von 114 (Statistisches Bundesamt 2018). Aus dieser Grundgesamtheit ist eine Stichprobe zu ziehen, wozu vorab

ein Untersuchungszeitraum definiert werden muss. An dieser Stelle sei auf Stamm (2013) verwiesen, der jährlich erscheint und in dem alle publizistischen Einheiten aufgeführt sind.

Viele Abschlussarbeiten in den Medienwissenschaften, der Publizistik aber auch in den Sozialwissenschaften untersuchen die Darstellung eines Themas in der Presse und beschränken sich dabei auf die Tagespresse. Somit werden in größerem Abstand erscheinende Zeitungen und Zeitschriften ausgeschlossen. Da eine Vollerhebung bei 114 Zeitungen nicht zu realisieren ist, muss aus dieser Grundgesamtheit der publizistischen Einheiten eine Stichprobe gezogen werden. Während bei Bevölkerungsumfragen die Personen entweder rein zufällig oder nach Quoten auswählt werden, ist dieses Vorgehen bei Zeitungen unangemessen. Diese unterscheiden sich insbesondere in ihrer Auflage bzw. Reichweite.

Je nach Erkenntnisinteresse wird meistens eine geschichtete Stichprobe (Häder/Häder, Kapitel 27 in diesem Band) gewählt, in der auch die Auflagenhöhe berücksichtigt wird. Stichprobenkriterien sind das Verbreitungsgebiet und die Vertriebsform, so dass aus nationalen überregionalen Abonnementszeitungen (z. B. Die Welt, FAZ, SZ und FR), Boulevardzeitungen (z. B. Bild, Abendpost Nachtausgabe, Express und Morgenpost Hamburg) und regionalen Abonnementzeitungen (z. B. Westdeutsche Allgemeine Zeitung (WAZ), Neue Osnabrücker Zeitung, Nürnberger Nachrichten) ausgewählt wird. Bei den ersten beiden Gruppen sind Vollerhebungen möglich, während man aus den regionalen Abonnementzeitungen eine Stichprobe ziehen sollte (dazu Früh 2001: 137–141 und Merten 1995: 283–292).

Die nächste Entscheidung betrifft die Analyseeinheiten, diese ergeben sich aus den Hypothesen und sind meistens die Artikel. Sind Zeitungen und Untersuchungszeitraum bestimmt, müssen die Artikel ausgewählt werden. Bei vielen Analysen werden im ersten Schritt alle zum Untersuchungsthema erschienenen Artikel ausgewählt. Muss aus diesen eine Stichprobe gezogen werden, dann bietet sich die *künstliche Woche* an, um Tageseffekte zu vermeiden. Eine künstliche Woche wird erreicht, in dem in der ersten Woche des Untersuchungszeitraums die Ausgabe vom Montag, in der zweiten Woche die Ausgabe vom Dienstag usw. genommen wird. Damit wird vermieden, dass eine Verzerrung durch Inhalte der Tageszeitungen stattfindet, so ist z. B. in Montagsausgaben der Sport überproportional vertreten, während bei den Samstagsausgaben die Unterhaltungsteile umfangreicher als an anderen Wochentagen sind. Wie wird über die Politikerinnen und Politiker berichtet, die nach der Bundestagswahl 2021 an den Sondierungs- und Koalitionsverhandlungen beteiligt waren? Der Zeitraum sollte sinnvollerweise die Tage zwischen Bundestagswahl (26. 9. 2021) und der Wahl des Bundeskanzlers (8. 12. 2021) umfassen. Die nächste Entscheidung ist, welche Medien untersucht werden sollen. Bei Zeitungen sind es bundesweit erscheinende Tageszeitungen wie Frankfurter Allgemeine Zeitung, Süddeutsche Zeitung, taz, Welt und Bild. Zusätzlich zu diesen Zeitungen kann man eine Auswahl von regional erscheinenden Tageszeitungen nehmen. Die Anzahl der berücksichtigten Zeitungen hängt von den zur Verfügung stehenden Ressourcen ab. Das gilt

auch für die Entscheidung, ob man alle in diesem Zeitraum erschienenen Artikel über das Thema oder nur eine Stichprobe analysieren möchte. Da der Zeitraum relativ kurz ist, macht eine Vollerhebung Sinn. Der Artikel ist die Analyseeinheit, in der das Vorkommen der im Forschungsinteresse stehenden Kategorien gezählt wird, also z. B. die Namen der Beteiligten, deren jeweilige Parteizugehörigkeit, ob und wie die jeweiligen Personen bewertet wurden und um welche Themen es in dem Artikel ging. Eigenschaften des Artikels wie Länge, Platzierung (auf welcher Seite), Datum der Veröffentlichung und Name der Zeitung sind für eine statistische Auswertung ebenfalls wichtig.

Ein weiteres Beispiel für die die Verwendung von Textdaten aus Zeitungen und Zeitschriften ist die Analyse von Heirats- und Kontaktanzeigen, mit denen ein Wertewandel zu untersucht werden kann. Dazu können überregionale Zeitungen verwendet werden, die möglichst über viele Jahre regelmäßig solche Anzeigen veröffentlichen, also Die Zeit sowie die Wochenendausgaben von FAZ, Welt und Süddeutscher Zeitung. Als Beispiel kann eine Studie von Klein (2013) genannt werden, der seinen Untersuchungszeitraum auf die Jahre 1950 bis 2005 festgelegte. Da eine Vollerhebung dieser Zeitungen zu aufwändig gewesen wäre, wurde eine Zufallsstichprobe gezogen. Von jeder Zeitung wurden pro Jahr vier Ausgaben ausgewählt, jeweils die erste zu Beginn eines neuen Vierteljahres. Von jeder dieser Ausgaben wurden jeweils 100 Anzeigen ausgewählt, waren es weniger, wurden alle in dieser Ausgabe abgedruckten Anzeigen in die Stichprobe aufgenommen. Des Weiteren wurden nur Anzeigen heterosexueller Menschen ausgewählt, da bis 1977 Homosexualität in Deutschland strafbar war. Anzeigen von mehreren Personen oder kommerzielle Anzeigen von Heiratsinstituten wurden in der Stichprobe ausgeschlossen, ebenso fremdsprachige Anzeigen. Jede Anzeige wurde aufgeteilt in Selbstbild – also wie beschreibt sich die inserierende Person selbst, in Fremdbild – also wie soll die gesuchte Person sein, in Beziehungsbild – wie soll die Beziehung gestaltet werden, und in die Kategorie sonstiges. Diese Aufteilung ermöglichte es, diese Inhalte über Zeit auszuwerten sowie geschlechtsspezifische Werte zu identifizieren (vgl. Klein 2013, 41–42, weitere Beispiele finden sich dort).

Für die Analyse des Untersuchungsmaterials bieten sich verschiedene Verfahren an:

- hypothesenprüfende: die sozialwissenschaftliche Inhaltsanalyse (content analysis)
- hypothesensuchende: die qualitative Datenanalyse (QDA – qualitative data analysis)

Bei hypothesenprüfenden Verfahren werden Kategorien gebildet und die resultierenden Häufigkeiten ausgezählt. Diese Daten können dann statistisch ausgewertet werden, z. B. nach dem Publikationszeitraum oder nach dem Geschlecht der Verfasser. Wichtig ist bei der Datenerhebung, dass auch externe Merkmale des Mediums wie Name der Zeitung, Erscheinungsdatum, Position des Artikels (Seite), Artikelart (z. B.

Aufmacher, Kommentar, Glosse, Reportage) berücksichtigt werden, damit diese in der Auswertung berücksichtigt werden können. Es wird ein Kategoriensystem und darauf basierend ein Codebuch entwickelt, so dass die Texte anhand von Beispielen codiert werden können. Die Kategorien sind standardisiert, alle haben einen numerischen Code, welche in ein Codesheet eingetragen werden. Für jeden Fall, also z. B. für jeden Artikel, gibt es ein separates Codesheet, welche auf einen elektronischen Datenträger übertragen und dann statistisch analysiert werden. Mit einem Pretest (Weichbold, Kapitel 28 in diesem Band) wird das Kategoriensystem auf seine Brauchbarkeit getestet, bevor die Codierung aller Artikel beginnt. Beim Kategoriensystem Sozwob von Dohrendorf gibt es z. B. die Kategorie „Verkehr". Immer wenn es in dem Leitkommentar um Verkehr geht oder Verkehr zumindest erwähnt wird, wird diese Kategorie in den Codesheet eingetragen (Dohrendorf 1990: 85). Kommen in einer Heirats- und Kontaktanzeige Namen von Dichtern, Musikern oder Malern vor, dann wird die Kategorie „Kultur" codiert; Kinder, Familie und Nestbau sind Indikatoren für die Kategorie „familiäre Orientierung". In der Kategorie „materielle Werte" sind Automarken, Wohneigentum wie z. B. eigenes Haus, aber auch Yachten oder Flugzeuge enthalten.

Die Reliabilität einer Codierung kann dadurch gewährleistet werden, dass mehrere Codierer den gleichen Text codieren und die Abweichungen analysiert werden – dies wird als Intercoderreliabilität bezeichnet. Durch den Codiervorgang können Gewöhnungs- und Lerneffekte bei den Codierern auftreten, eine derartige Intracodierreliabilität kann gemessen werden, in dem die gleichen Texte zu Anfang und zum Ende des Projektes von denselben Codierern codieren werden. Der Codiervorgang selbst nimmt viel Zeit in Anspruch, die bei der Projektplanung berücksichtigt werden muss.

Statt einer konventionellen kann man auch eine computerunterstützte Inhaltsanalyse durchführen, die auf syntaktischer Ebene arbeitet. In diesem Fall werden Suchbegriffe – das können ganze Wörter, Wortteile oder Kombinationen davon sein – gezählt. Auf der Basis von Wörterlisten wird dann ein Kategoriensystem entwickelt, was meistens mit entsprechender Software erfolgt. Bei dieser Vorgehensweise kann das Kategoriensystem jederzeit und mit relativ wenig Aufwand geändert werden. Die Bedeutung der Wörter wird in Form von Kategorien berücksichtigt: ähnliche Suchbegriffe werden in einer Kategorie mit einem numerischen Code zusammengefasst und entsprechend ausgezählt. Je öfter eine Kategorie genannt wird, desto wichtiger ist sie. Allerdings treten semantische Probleme wie Mehrdeutigkeit und Negation auf, die bei einer konventionellen Inhaltsanalyse durch die Sprachkompetenz der Codierer gelöst werden. Mehrdeutige Suchbegriffe können dadurch codiert werden, dass diese schon im Kategoriensystem als solche markiert werden. Beim Auftreten im Text stoppt die Codierung, und man kann dann selbst entscheiden, ob und wenn ja, mit welcher Kategorie codiert wird (interaktive Codierung). Alternativ können auch Suchbegriffe definiert werden, die ein oder mehrere Wörter zusätzlich enthalten und so die Mehrdeutigkeit auflösen. Negierte Suchbegriffe können im Deutschen

an bestimmten Wort(teilen) erkannt werden, entweder stehen Wörter wie nicht oder kein direkt vor oder direkt nach dem Suchbegriff, oder die Vorsilbe un steht vor dem Suchbegriff oder los (z. B. arbeitslos) steht direkt dahinter. Auch doppelte Negation können identifiziert werden, beim Suchbegriff arbeit, die Textstelle nicht arbeitslos. Hier sind die Indikatoren für Negation das Wort nicht vor dem Suchbegriff arbeit und los direkt am Suchbegriff angehängt (vgl. Klein 2013: 100–104).

Die Arbeitstechniken von computergestützten Textanalysen sind unterschiedlich (Klein 2010). Bei den hypothesenprüfenden *quantitativen* Verfahren wird ein Kategoriensystem erstellt, dessen Suchbegriffe mit dem Ziel der statistischen Analyse ausgezählt werden, dabei können auch Zusammenhänge zwischen den Variablen des Kategoriensystems und den externen Variablen untersucht werden.

Bei der *qualitativen* Datenanalyse (Taddicken, Kapitel 107 in diesem Band) werden die wichtigen Textstücke am Bildschirm markiert, und so entsteht sukzessive ein Kategoriensystem. Auch hier können externe Variablen definiert und mit den Kategorien verknüpft werden. Entsprechende Software wurde seit den 1980er Jahren immer weiter entwickelt und ermöglicht diese Art von Analysen.

Textmining (Manderscheid, Kapitel 121 in diesem Band) ist eine Analyseform, mit der ohne ein Kategoriensystem und ohne manuelle Codierung Texte untersucht werden. Dazu wird das gemeinsame Auftreten von Wörtern gezählt und in Beziehung gesetzt, daraus kann dann ein Wortfeld entstehen. Bestimmte Wortarten wie Artikel, Präpositionen oder Konjunktionen werden dabei nicht berücksichtigt.

Für die Inhaltsanalyse selbst stehen gute Lehrbücher zur Verfügung. Einen umfangreichen Überblick über inhaltsanalytische Verfahren gibt Merten (1995), Früh (2001) zeigt detailliert an vielen Beispielen die Anwendung. Krippendorff (2012) gilt als Standardwerk im englischsprachigen Raum.

Literatur

Dohrendorf, Rüdiger (1990): Zum publizistischen Profil der „Frankfurter Allgemeinen Zeitung“: computerunterstützte Inhaltsanalyse von Kommentaren der FAZ. Frankfurt am Main: Peter Lang Verlag

Früh, Werner (2001): Inhaltsanalyse. Theorie und Praxis. 5., überarb. Auflage. Konstanz: UVK

Klein, Harald/Giegler, Helmut (1994): Correspondence Analysis of Text Data with INTEXT/PC. In: Greenacre, Michael/Blasius, Jörg (Hg.): Correspondence Analysis in the Social Sciences. London: Academic Press, 283–301

Klein, Harald (2010): Inhaltsanalyse. In: Atteslander, Peter (Hg.): Methoden der empirischen Sozialforschung, 13. Auflage. Berlin: Schmidt Verlag, 195–224

Klein, Harald (2013): Computerunterstützte Textanalysen mit TextQuest. Eine Einführung in Methoden und Arbeitstechniken. München und Mering: Rainer Hampp Verlag

Krippendorff, Klaus (2012): Content Analysis, An Introduction to Its Methodology, 3. Auflage. Thousand Oaks, CA: Sage

Merten, Klaus (1995): Inhaltsanalyse. Einführung in Theorie, Methode und Praxis, 2. Auflage. Opladen: Westdeutscher Verlag

Pürer, Heinz/Raabe, Johannes (1996): Medien in Deutschland. Band 1: Presse, 2. Auflage. Konstanz: UVK

Stamm, Willy (2013): STAMM Leitfaden durch Presse und Werbung 2013: Presse- und Medienhandbuch. Essen: STAMM Verlag

Statistisches Bundesamt (2018): Anzahl von Tageszeitungen in Deutschland. URL: https://de.statista.com/statistik/daten/studie/36376/umfrage/anzahl-von-tageszeitungen-in-deutschland-seit-1965/

Weitzman, Eben/Miles, Matthew B. (1995): Computer Program for Qualitative Data Analysis. A Software Sourcebook. Thousand Oaks: Sage

Weblinks (aufgerufen am 20. 11. 2021)

http://ww.textanalysis.info gibt seit 1999 einen ständig aktualisierten Überblick über Textanalysesoftware. Dort werden die Programme kurz beschrieben, es gibt Links zu den Handbüchern und zu kostenlosen Test- oder Demoversionen, so dass man ausprobieren kann, welche Software am besten für das eigene Projekt geeignet ist.

http://de.wikipedia.org/wiki/Publizistische_Einheit

Harald Klein ist Gründer und Inhaber der Firma Social Science Consulting und war an verschiedenen Universitäten tätig. *Ausgewählte Publikationen:* Computerunterstützte Textanalysen mit TextQuest. Ein Einführung in Methoden und Arbeitstechniken. München/Mering: Hampp (2013); Correspondence Analysis of Text Data with INTEXT/PC, in: Michael Greenacre und Jörg Blasius (Hg.): Correspondence Analysis in the Social Sciences. London: Academic Press (zusammen mit Helmut Giegler, 1994). *Kontaktadresse:* hklein@textquest.de.