Computerunterstützte Textanalysen mit TextQuest

SOZIALWISSENSCHAFTLICHE
FORSCHUNGSMETHODEN

herausgegeben von
Wenzel Matiaske, Martin Spieß,
Ingwer Borg, Claudia Fantapié-Altobelli, Holger Hinz,
Uwe Jirjahn, Bernhard Kittel, Manfred Kraft,
Stefan Liebig, Rainer Oesterreich, Jost Reinecke,
Kai-Uwe Schnapp, Rainer Schnell, Peter Sedlmeier,
Winfried Seidel, Gerhard Tutz, Joachim Wagner

Band 6

Harald Klein

Computerunterstützte Textanalysen mit TextQuest

Eine Einführung in
Methoden und Arbeitstechniken

Rainer Hampp Verlag München und Mering 2013

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

ISBN 978-3-86618-831-0 (print)
ISBN 978-3-86618-931-7 (e-book)
SOZIALWISSENSCHAFTLICHE FORSCHUNGSMETHODEN: ISSN 1869-7151
DOI 10.1688/9783866189317
1. Auflage, 2013

 Rainer Hampp Verlag München und Mering
Marktplatz 5 D – 86415 Mering
www.Hampp-Verlag.de

∞ *Dieses Buch ist auf säurefreiem und chlorfrei gebleichtem Papier gedruckt.*

Liebe Leserinnen und Leser!
Wir wollen Ihnen ein gutes Buch liefern. Wenn Sie aus irgendwelchen Gründen nicht zufrieden sind, wenden Sie sich bitte an uns.

Kapitel 1
Einleitung

1.1 Für wen dieses Buch ist – und für wen nicht

Das Buch richtet sich an alle wissenschaftlichen Disziplinen, die Texte analysieren – Wirtschafts- und Sozialwissenschaften, Literatur- und Sprachwissenschaften, Rechtswissenschaft oder Medizin. Anwendungsgebiete von TEXTQUEST sind Texte aller Art, seien es Fernsehnachrichten, um senderspezifischen Nachrichtenfaktoren herauszufinden oder dem Wertewandel mit Heirats- und Kontaktanzeigen auf die Spur zu kommen. Auch Lesbarkeitsanalysen von Parteiprogrammen und medizinischen Beipackzetteln sind möglich.

In diesem Buch liegt der Schwerpunkt auf Arbeitstechniken, die für die Planung und Durchführung von computerunterstützten Inhaltsanalysen unerlässlich sind. Dazu gehören die Beschaffung und die Aufbereitung von Texten, die Operationalisierung von Hypothesen und deren adäquate Auswertung. Grundlegende Kenntnisse der Methoden der empirischen Sozialforschung werden vorausgesetzt, insbesondere Hypothesenbildung und deren Implikationen, Reliabilität und Valididät sowie Probleme der Messung. Theoretische Grundlagen werden kurz vorgestellt, weiterführende Literatur findet sich in den entsprechenden Kapiteln, insbesondere haben Merten (1995) und Krippendorff (2004) die theoretische Fundierung detailliert dargestellt.

Computerunterstützte Verfahren erfordern Software, für dieses Buch wurde der Schwerpunkt auf TEXTQUEST gelegt. Programme, die keine oder nur eingeschränkt selbst entwickelte Kategoriensysteme verarbeiten können wie Diction, LIWC oder PCAD werden in diesem Buch nicht berücksichtigt. Auf http://www.textanalysis.info sind weitere Informationen über Textanalysesoftware beschrieben.

TEXTQUEST (und sein Vorläufer INTEXT) wird seit 1988 entwickelt. Schwerpunkt sind die computerunterstützte Inhaltsanalyse und die Lesbarkeitsanalyse mit Lesbarkeitsformeln.

In diesem Buch wird der Begriff Menu verwendet und nicht Menü. Menu ist eine Speisekarte, aus der man wählen kann; dies ist bei der Beschreibung

von Benutzeroberflächen gemeint. Menu mit Menü zu übersetzen ist falsch, hat sich leider aber eingebürgert.

1.2 Die Beispieldaten

Als Beispieldatensatz werden in diesem Buch Heirats- und Kontaktanzeigen verwendet. An dieser Stelle möchte ich der Universität Osnabrück Dank sagen, die die Datenerfassung aus ihrem Forschungsförderungsfond zum großen Teil finanzierte. Die Datenerfassung erfolgte mit studentischen Hilfskräften im Zeitungsarchiv der Westfälischen Wilhelms-Universität Münster, deren guter Service und Entgegenkommen an dieser Stelle großer Dank gebührt.

Mit den Heirats- und Kontaktanzeigen soll der Wertewandel in Deutschland untersucht werden. Gerade diese Anzeigen enthalten die Werte, die für die Inserierenden für sein persönliches Leben wichtig sind.

Die Heirats- und Kontaktanzeigen stammen aus dem Zeitraum von 1950 bis 2005 aus den Zeitungen *Frankfurter Allgemeine Zeitung, Westdeutsche Allgemeine Zeitung, Süddeutsche Zeitung* und *Die Zeit*. Von jeder Zeitung wurden je Jahr 100 Stück zu vier verschiedenen Zeitpunkten innerhalb eines Jahres ausgewählt. Da in einigen Zeitungen nicht jedes Jahr 100 Anzeigen erschienen, wurden 21451 Heiratsanzeigen analysiert.

Die Anzeigentexte wurden geteilt in:

- das Selbstbild: wie sich die inserierende Person beschreibt, Code 1
- das Fremdbild: wie die gesuchte Person sein soll, Code 2
  Beispiel: Partner mit Adjektiv (netter Partner)
- das Beziehungsbild: wie eine Beziehung aussehen soll, z.B. gemeinsame Unternehmungen oder Aufbau einer Beziehung. Code: 3
  Beispiel: wenn eine Ehe spezifiziert wird, z.B. Idealehe oder Neigungsehe
- sonstiges: Angaben wie Bildzuschrift, Telefonnummern, Vermittlung von Eltern oder Freunden. Code 4

Die Anzeigentexte wurden daher mit entsprechenden Steuerkommandos versehen, dabei wurden die Texte auch normiert. Ein Beispiel sind die Postleitzahlen, die sich im Umtersuchungszeitraum zweimal geändert haben. Diese wurden durch die entsprechenden Ortsangaben ersetzt, damit diese über die Zeit gesehen vergleichbar sind.

In diesem Buch wird beschrieben, wie die Texte aufbereitet wurden, welche Problem dabei auftraten und wie sie gelöst wurden.

1.3 Historisches

Die Entwicklung des sozialwissenschaftlichen Verfahrens der Inhaltsanalyse wird von Merten (1995: 35-46) in mehrere Phasen eingeteilt, die sich überlappen:

- die Phase der Intuition dauerte bis 1900. Der Schluss von der Beschaffenheit manifester Inhalte auf nichtmanifeste Umstände ist charakteristisch.
- die quantitativ-deskriptiven Phase setzt Merten vom 7. Jahrhundert bis 1926 an. Bei den Analysen handelt es sich um Häufigkeiten von Worten, also um Textanalysen auf der syntaktischen Ebene.
- die Entwicklung der Inhaltsanalyse zu einem eigenständigen Verfahren der Datenerhebung ist von 1926 – 1941 ziemlich kurz. Sie ist durch die Entwicklung der Massenpresse inspiriert. Durch das Aufkommen weiterer Massenmedien wie Film, Radio und Fernsehen entwickelte sich die Werbewirkungsforschung und somit das Datenerhebungsinstrument zum eigenständigen Datenerhebungsinstrument (Merten 1995: 38-41; Krippendorff 2004: 3-17).
- die interdisziplinären Erweiterung dauert laut Merten (1995: 41-44) von 1941 - 1967 und begann 1941 auf einer Konferenz in Chicago, auf der unter anderem der Begriff der Inhaltsanalyse – content analysis – geprägt wurde.
- die Phase theoretisch-methodischen Fundierung dauert seit 1967 an und beinhaltet den Aufstieg der computerunterstützten Inhaltsanalyse, die auf einer Konferenz in Annenberg (PA, USA) begann.

7 Jahre später auf einer Konferenz in Pisa stellte Alexander Deichsel erste Ergebnisse einer Analyse von Zeitungsschlagzeilen vor. Ein Grund für die Weiterentwicklung computerunterstützter Verfahren ist darin zu sehen, dass im Verlaufe der Zeit die Hard- und Software immer leistungsfähiger wurde.

Die hier diskutierten Programme haben auch ihre Geschichte, sie begann mit der Entwicklung des General Inquirer 1961 (Stone u.a. 1966: IX). 1969 wurde der General Inquirer in Europa vorgestellt, dies führte zu Neuentwicklungen in vielen Ländern, die teilweise bis heute andauern. Allerdings wurde viele Projekte nach einigen Jahren eingestellt. Oft lag der Grund darin, dass die Entwickler die rasante Entwicklung auf dem Computerbereich nicht mitmachen konnten. Viele interessante und vielversprechende Entwicklungen wurden deshalb nicht fortgeführt. Einen Überblick und weitere Hintergründe geben mit deutschem Schwerpunkt Klein (1996: 25-38) und mit amerikanischem Schwerpunkt Diefenbach (2001: 13-32).

Viele der heutigen Programme kann man als Varianten des Urvaters *General Inquirer* sehen, oft bieten sie eine ähnliche Funktionalität an. Der General Inquirer kann unter anderem – allerdings nur für die englische Sprache – zwischen mehreren Bedeutungen eines Wortes unterscheiden, in dem der Kontext des Wortes für die Unterscheidung der Bedeutung herangezogen wird. Ebenso ist eine Formulierung von komplexen Suchbegriffen, die kodiert

werden sollen, möglich. Auch KWICs (**K**ey-**W**ord-**I**n-**C**ontext) können erstellt werden. Unter einem KWIC versteht man den einzeiligen Ausdruck eines Suchbegriffs im Kontext. Die erste Version des General Inquirer stand 1966 zur Verfügung, allerdings lief sie nur auf einer bestimmten Hardware. 1987 wurde sie von ZUMA (Zentrum für Umfragen, Methoden und Analysen, heute GESIS Manheim) in Mannheim überarbeitet und neu dokumentiert (ZUMA-Nachrichten 20 (1987): 32-36, Züll u. a. 1989). Der General Inquirer ist in PL/1 geschrieben, eine Programmiersprache, die im wesentlichen auf IBM-Großrechner beschränkt ist[1].

Philip Stone stellte 2001 eine in Java geschriebene Version namens New Inquirer vor, die Excel-Tabellen benutzt, in der die Kategoriensysteme abgespeichert sind. Der Vorteil des New Inquirer ist, dass dieses Programm betriebssystemunabhängig ist, da es in Java geschrieben ist. Für den General Inquirer sind zwei Kategoriensysteme Standard, dass Lasswell-Value-Dictionary und das Harvard-Dictionary. Beide Kategoriensysteme sind für mehrere Sprachen und auch für mehrere Programmpakete verfügbar und haben eine allgemeingültige Zielsetzung. Inwieweit nach seinem Tod im Januar 2006 das Projekt weitergeführt wird, ist zurzeit unklar.

Unter http://www.wjh.harvard.edu/ inquirer/server_blognote.html können weitere Details bis zum März 2004 einschliesslich dem Blog des General Inquirers entnommen werden.

Eines der ersten und heute noch existierenden Programme ist Textpack. 1969 begannen Dieter Fuchs, Hans-Dieter Klingemann, Jürgen Höhe und Klaus Radermacher im Rahmen eines DFG-Projektes die Entwicklung von Textpack. Die Versionen I und II wurden am Zentralarchiv für empirische Sozialforschung in Köln (heute GESIS Köln) entwickelt, die danach folgenden Versionen III bis V im Rahmen der Grundlagenforschung ab 1974 beim Zentrum für Methoden, Methoden und Analysen (ZUMA)in Mannheim (vgl. Klingemann 1984: 17). Da Textpack nur auf Großrechnern der Typen IBM und Siemens ablauffähig war, beschloss ZUMA eine Neuprogrammierung in Fortran 77 namens Textpack V. 1987 wurde eine PC-Version für MS-DOS vorgestellt. Textpack V enthält nicht alle Möglichkeiten, die die vorherige Version IV bietet[2]. Für Textpack werden regelmäßig Workshops bei GESIS Mannheim angeboten, Menus und Handbuch wahlweise in Englisch oder Spanisch. Deutsche Menus gibt es nicht, wohl aber eine Kurzbeschreibung in Deutsch. Die zu analysierenden Texte können in Sprachen mit lateinischem Alphabet vorliegen. In den letzten Jahren hat es keine Weiterentwicklung gegeben.

---

[1] Seit 1995 gibt es eine Implementation für das IBM-Betriebssystem OS/2, dieses Betriebssystem wird heute (2013) kaum noch genutzt.

[2] Die Version V ist bei der Definition von Suchbegriffen teils leistungsfähiger als die Version IV, weil auch Textteile, die am Anfang eines Wortes stehen, kodiert werden können, teils weniger leistungsfähig, weil Kombinationen von Einzelworten nicht mehr kodiert werden können.

TEXTQUEST wurde vom Autor bis 1999 unter dem Namen INTEXT (INhaltsanalyse von TEXTen) entwickelt, 1981-1987 auf einem IBM-Großrechner unter Verwendung der Programmiersprache PL/1, danach unter MS-DOS und Verwendung der Programmiersprache C. 1999 erfolgte die Umbenennung in TEXTQUEST, weil es ein Intext for Windows schon gab. Seitdem erfolgt die Programmierung in C++ unter Verwendung von wxWidgets. In Version 3 wurde das Modul zum Vergleichen von Vokabularen (z.B. Wörterlisten) erweitert, wobei auch mehr als zwei Vokabulare gleichzeitig verglichen werden können. Ebenso wurde ein Kategorienmanager eingebaut, der das bequeme Entwickeln von Kategoriensystemen erlaubt. Seit Version 4 läuft TEXTQUEST auch unter Apples Mac OS-X und kann Texte verarbeiten, die in Latin-1 oder UTF-8 kodiert wurden; das ermöglicht die Verwendung von vielen Sprachen mit lateinischem Alphabet inklusive Buchstaben mit Akzenten und/oder Diakritika.

Das Programm gibt es mit englischsprachigem Handbuch. Menus und Hilfetexte gibt es sowohl in Englisch als auch in Deutsch.

WordStat wurde von Normand Peladeau, Montreal, Kanada, entwickelt. Es ist ein Add-on für seine Programme SimStat oder QDA-Miner, allein ist WordStat nicht abläuffähig. Dieses Design hat den Vorteil, dass WordStat zusammen mit SimStat das Datenmanagement und die statische Auswertung erheblich erleichtert. Menus und Dokumentation sind in Englisch, Französisch oder Spanisch. Im Gegensatz zu den anderen Programmen bietet WordStat auch visuelle Auswertungen an.

1.4 Definitionen

In diesem Kapitel werden die wichtigen Begriffe und ihre teilweise verschiedenen Definitionen vorgestellt und diskutiert. Wer weiter ins Detail gehen möchte, sei auf Merten (1995) und Krippendorff (2004) verwiesen.

1.4.1 Inhaltsanalyse

Die Definitionen zum Begriff der Inhaltsanalyse sind völlig unterschiedlich und stellen verschiedene Aspekte in den Mittelpunkt des Forschungsinteresses. Ausführliche Diskussionen finden sich in Merten (1995: 14-35), Krippendorff (2004: 18-21) und Klein (1996: 12-14).

Im folgenden soll Mertens Definition als Arbeitsgrundlage dienen:

> „Inhaltsanalyse ist eine Methode zur Erhebung sozialer Wirklichkeit, bei der von Merkmalen eines manifesten Textes auf Merkmale eines nicht manifesten Kontextes geschlossen wird.“ (Merten 1995: 15-16)

Der Untersuchungsgegenstand ist die *soziale Wirklichkeit*, die Kommunikationsinhalte müssen aber in manifesten Text überführbar sein (vgl. Merten 1995: 16).

Kernpunkt jeder Inhaltsanalyse ist die Bildung von Kategorien, die ihrerseits aus theoretischen Annahmen abgeleitet werden. Die Bildung von Kategorien ist das Kernstück einer Inhaltsanalyse.

> „Die Inhaltsanalyse – in ihrer klassischen Form – ist ein weitgehend nichtreaktives Verfahren zur Gewinnung von (vorwiegend symbolischen) Daten und zur Verarbeitung und Analyse solcher Daten mit Hilfe von Kategorien, die ihrerseits eng mit theoretischen Annahmen über einen Phänomenbereich verknüpft sind." (Fischer 1982: 179)

Fischer bezeichnet die Inhaltsanalyse explizit als ein Datenerhebungsverfahren und spezifiziert die Methode, verwendet allerdings Begriffe wie „weitgehend nichtreaktiv", „vorwiegend symbolisch" und „Phänomenbereich", die unpräzise sind und die er nicht erklärt.

Berelson wies auf die Schlüsselposition der Kategorien hin:

> „Da die Kategorien die Substanz der Untersuchung enthalten, kann eine Inhaltsanalyse nicht besser sein als ihre Kategorien." (Berelson 1952: 147)

Kategorien nehmen im Verfahren der Inhaltsanalyse eine Schlüsselstellung ein, deren Position Werner Früh verdeutlicht:

> „Der Sinn jeder Inhaltsanalyse besteht letztlich darin, unter einer bestimmten forschungsleitenden Perspektive Komplexität zu reduzieren. Textmengen werden hinsichtlich theoretisch interessierender Merkmale klassifizierend beschrieben. ... Nach angegebenen Kriterien werden einige Mitteilungsmerkmale als untereinander ähnlich betrachtet und einer bestimmten Merkmalsklasse bzw. einem Merkmalstypus zugeordnet, den man in der Inhaltsanalyse 'Kategorie' nennt." (Früh 2001: 39-40).

Noch deutlicher beschreiben Lisch und Kriz, dass bei der Kategorienbildung eine Informationsreduktion stattfindet:

> „Der Sinn der Kategorienbildung besteht also im wesentlichen in der Informationsreduktion, indem sich der Inhaltsanalytiker nach dem Kodierungsprozeß, in dem die ursprünglichen Textbestandteile den Kategorien zugeordnet werden, auf die Behandlung einiger weniger Kategorien anstelle einer Vielzahl von kleineren Textelementen beschränkt." (Lisch/Kriz 1978: 70)

Sowohl Früh als auch Lisch und Kriz messen der Kategorienbildung und dem Kodierungsprozess eine weitreichende Bedeutung bei. Deshalb ist es wichtig, dass Kategorien nach den im empirischen Forschungsprozess relevanten Kriterien gebildet werden. Sie müssen vom Erkenntnisinteresse geleitet und in Hypothesen fixiert sein, die verifiziert oder falsifiziert werden können. Kategorien werden in ein oder mehrere Variablen (Merkmale) gefasst, die verschiedene Ausprägungen haben.

Die Gesamtheit der Kategorien einer inhaltsanalytischen Untersuchung wird im folgenden als Kategoriensystem bezeichnet.

Die folgenden Kriterien muss ein Kategoriensystem erfüllen: (vgl. Holsti 1969: 95 und Merten 1995: 98-105)

- Das Kategoriensystem muss aus den Untersuchungshypothesen theoretisch abgeleitet sein.
- Die Kategorien eines Kategoriensystem müssen voneinander unabhängig sein (d.h. sie dürfen nicht miteinander korrelieren). Das ist besonders für die statistische Auswertung wichtig.
- Die Ausprägungen jeder Kategorie müssen vollständig sein.
- Die Ausprägungen jeder Kategorie müssen wechselseitig exklusiv sein, sie dürfen sich nicht überschneiden und müssen trennscharf sein.
- Die Ausprägungen jeder Kategorie müssen nach einer Dimension ausgerichtet sein (einheitliches Klassifikationsprinzip).
- Jede Kategorie und ihre Ausprägungen müssen eindeutig definiert sein.

Mit dem Kategoriensystem werden die Regeln der Kodierung festgelegt. Damit werden Merkmale des Kommunikationsinhaltes in numerische Daten überführt. Dieser Vorgang heisst Verschlüsselung oder Kodierung. Dabei treten Probleme auf, die Reliabilität und Validität der Inhaltsanalyse beeinflussen. Mit diesen Kriterien wird die Güte einer Inhaltsanalyse beurteilt. Beide Kriterien müssen möglichst gut erfüllt sein.

Unter Validität (Gültigkeit) ist zu verstehen, ob mit der Inhaltsanalyse wirklich das gemessen wird, *was* gemessen werden soll, ob also das Messinstrument für die Überprüfung der Hypothesen geeignet ist. Die Validität hängt davon ab, wie präzise die Kategorien des Kategoriensystems definiert sind und ob diese Operationalisierung plausibel (face-validity) und vor allen Dingen auch brauchbar ist (vgl. Früh 2001: 183-186).

Die Reliabilität meint die Verlässlichkeit der Messung, also ob bei gleichem Analysematerial und gleichem Kategoriensystem die Ergebnisse gleich sind. Sie ist ein Gütekriterium für den Messvorgang, also *wie* gemessen wird. Dabei werden in der Inhaltsanalyse zwei Arten von Reliabilität unterschieden:

Interkoderreliabilität: Darunter werden die Unterschiede zwischen mindestens zwei verschiedenen Kodierern verstanden (vgl. Früh 2001: 177). Es gibt für die Messung dieser Reliabilität verschiedene Koeffizienten für die Überprüfung von zwei (Scott 1955) oder mehreren Kodierern (Craig 1981, Krippendorff 2004: 221-235). Die Werte liegen zwischen 0 und 1, sie sollten möglichst dicht an 1 liegen. Gute Kategorien haben einen Reliabilitätskoeffizienten, dessen Wert über 0,7 liegt. Die Interkoderreliabilität hängt von der Anzahl der Ausprägungen eines Merkmals, der Kodiererschulung, der Sorgfalt der Kodierung und der Güte des Kategoriensystems und seiner Definitionen ab (vgl. dazu Merten 1995: 304-310, Früh 2001: 115, 177-183).

Intrakoderreliabilität: werden die Unterschiede zwischen derselben kodierenden Person genannt (vgl. Früh 2001: 115). Sie wird von den gleichen Merkmalen wie die Interkoderreliabilität beeinflusst, dazu kommen noch Lerneffekte und eventuell Änderungen des Kategoriensystems, die während der Analyse notwendig wurden. Zur Messung wird der gleiche Text mit dem gleichen Kategoriensystem der gleichen Person zweimal (oder

noch öfter) in einem zeitlichen Abstand vorgelegt, idealerweise zu Beginn und zum Ende der Untersuchung.

Inhaltsanalysen müssen sowohl valide als auch reliabel sein. Beide Kriterien hängen zusammen, denn für die Güte eines Instrumentes

„...ist hohe Zuverlässigkeit auch notwendige Voraussetzung für Gültigkeit. ... Umgekehrt allerdings muß auch ein sehr zuverlässiges Instrument noch lange nicht gültig sein." (Kriz 1978: 85).

Damit ist gemeint, dass fehlerhafte Messinstrumente nicht mehr das messen, was gemessen werden soll. Da sie nicht mehr reliabel sind, können sie auch nicht mehr valide sein. Ist ein Messinstrument reliabel, muss es nicht notwendigerweise auch valide sein. Man kann Merkmale perfekt reliabel messen, die zur Überprüfung der aufgestellten Hypothesen ungeeignet sind. Reliabilität ist eine Voraussetzung für Validität, aber Validität keine Voraussetzung für Reliabilität.

Quantitative und qualitative Aspekte sind bei der Inhaltsanalyse nicht trennbar, denn im empirischen Forschungsprozess steht die Überprüfung von Hypothesen oder Theorien – ob a-priori aufgrund theoretischer Vorüberlegungen oder a-posteriori bei der Auseinandersetzung mit dem Untersuchungsmaterial aufgestellt – vor der Durchführung einer Inhaltsanalyse, und sie beeinflussen sie – explizit oder implizit – bei der Bildung des Kategoriensystems:

„Die Auszählung der Häufigkeiten einer Kategorie setzt qualitative Vorleistungen bei der Festlegung eben dieser Kategorie voraus." (Huber 1989: 33)

Die Bemerkung Fischers, dass

„ein Inhalt, der quantifiziert wird, immer auch schon *qualifiziert* sein muß – es muß etwas von bestimmter Qualität schon da sein, um überhaupt gezählt werden zu können" (Fischer 1982: 181, Hervorhebung im Original)

zeigt, dass eine strikte Unterscheidung in rein quantitative oder rein qualitative Inhaltsanalysen nicht sinnvoll erscheint. In der Forschungspraxis ist es meist so, dass die qualitative Inhaltsanalyse als Verfahren zur Generierung von Hypothesen dient, während die quantitative Inhaltsanalyse ein Instrument zur Überprüfung von Hypothesen ist (vgl. dazu Spöhring 1989: 9, 41-49).

1.4.2 Computerunterstützte Inhaltsanalyse

Die computerunterstützte Inhaltsanalyse unterscheidet sich von der konventionellen dadurch, dass der Kodierungsprozess nicht von Menschen, sondern von einem Computer mittels eines EDV-Programmes (Software) durchgeführt wird:

„Unter dem Begriff 'Computerunterstützte Inhaltsanalyse' versteht man Verfahren, die die Analyse und Vercodung der zu untersuchenden Inhalte mittels EDV-Programmen vornehmen. Voraussetzung für die Anwendung derartiger Verfahren ist, dass das Untersuchungsmaterial vertextet, d.h. in schriftlicher Form vorliegt" (Hörschinger 1985: 18)

Software wird vom Computer ausgeführt. Voraussetzungen dafür sind ein digitalisierter Text und ein digitalisiertes Kategoriensystem. Digitalisiert bedeutet, dass die Daten als Datei(en) auf einem Datenträger (z.B. Festplatte) gespeichert sind. Die verschiedenen Digitalisierungsmöglichkeiten werden im Kapitel 2 auf Seite 17 beschrieben.

Basis eines Kategoriensystems ist ein Verzeichnis aller im Text vorkommenden Zeichenketten – das sind alle Zeichen, die zwischen zwei Leerzeichen stehen (meist Wörter) – und ihrer Häufigkeit. Das Kategoriensystem besteht aus mindestens einer Kategorie mit wenigstens einem Suchbegriff (meistens einem Wort, eine komplette Beschreibung erfolgt in Kapitel 6.1 auf Seite 88) operationalisiert. Zu jedem Suchbegriff gehört ein numerischer Code, der immer dann vergeben wird, wenn der Suchbegriff im Text gefunden wurde. Die einzelnen Schritte dieses Verfahrens unterscheiden sich erheblich von der konventionellen Inhaltsanalyse (vgl. dazu auch Deichsel 1975: 47-48, Klein 1996: 20, Klein 2010: 218-221), wie Abbildung 1.1 zeigt.

Abb. 1.1: Generelle Unterschiede der Arten von Inhaltsanalyse

konventionell	computerunterstützt
Text durch Lesen eines kleinen Teils kennenlernen	Text digitalisieren und dabei kennenlernen
entfällt	Wörterliste als Basis des Kategoriensystems erzeugen
Kategoriensystem durch Definitionen und Beispiele aufbauen	Kategoriensystem durch vollständige Aufzählung der Suchbegriffe aufbauen
Kodiererschulung durchführen	entfällt
Pretest durchführen	Pretest durchführen
Kodierung durchführen	Kodierung durchführen
Daten auf Kodierblatt eintragen und von dort in Computer eingeben	entfällt
Daten bereinigen	entfällt
Daten auswerten	Daten auswerten

Insbesondere der Zeitaufwand für die einzelnen Schritte differiert erheblich. Während bei der computerunterstützten Inhaltsanalyse (cui) das Umsetzen auf Datenträger (Digitalisierung) der zeitintensivste Teil ist, ist es bei der konventionellen Inhaltsanalyse die Kodierung. Die Konstruktion eines Kategoriensystems nimmt bei beiden Formen etwa die gleiche Zeit in Anspruch. Bei der konventionellen Inhaltsanalyse sind die Übertragung der Kodierergeb-

nisse – zuerst auf Kodierblätter (Codesheets), dann davon auf Datenträger – zwei zeitraubende und fehleranfällige Arbeitsschritte, die bei der computerunterstützten Inhaltsanalyse entfallen, weil beim Kodieren der Rohdatensatz so gespeichert wird, dass er danach ohne großen Aufwand statistisch ausgewertet werden kann. Überprüfungen der Konsistenz der Codes, Filterfragen und der Reliabilität treten nach der Kodierung nicht auf. Probleme der Validität müssen bei der Konstruktion des Kategoriensystems berücksichtigt werden. Insbesondere muss beachtet werden, dass die Ergebnisse als Zähler für die Häufigkeit der einzelnen Codes in der Analyseeinheit vorliegen.

Bei der Durchführung einer computerunterstützten Inhaltsanalyse ergeben sich weitere Unterschiede in den Rahmenbedingungen, die für dieses Verfahren sprechen (Klein 1996: 21):

Abb. 1.2: Unterschiede der Rahmenbedingungen von Inhaltsanalyse

konventionell	computerunterstützt
bearbeitbare Textmenge von Zahl der Kodierer abhängig	bearbeitbare Textmenge von Datenträgerkapazität abhängig
bedingt komplexe Kategoriensysteme verwendbar	komplexe Kategoriensysteme verwendbar
Kodierung innerhalb von Wochen	Kodierung innerhalb von Minuten
meist nur eine Kodierung	beliebig viele Kodierungen
Kategoriensystem ist während der Kodierung kaum änderbar	Kategoriensystem kann leicht geändert und erweitert werden

Die Erfassung eines Textes am Bildschirm ist billiger als Kodierung, Datenerfassung und Datenbereinigung desselben Textes. Daher sind bei einer konventionellen Inhaltsanalyse mehrfache Kodierungen und Erweiterungen selten. Letztere sind nur dann problemlos, wenn das bereits kodierte Material nicht betroffen ist. Die Anzahl der Suchbegriffe eines Kategoriensystems wird bei der heutigen (2013) Computerhardware praktisch nicht beschränkt. Beliebig viele Kodierungen ermöglichen außerdem eine iterative Entwicklung des Kategoriensystems.

Jede Form der Inhaltsanalyse hat auch ihre Probleme, wie Abbildung 1.3 (Klein 1996: 22) zeigt.

Die wichtigste Einschränkung der computerunterstützten Inhaltsanalyse besteht darin, dass nur Hypothesen überprüfbar sind, die sich mit syntaktischen Kriterien – das sind in der Regel Einzelworte – operationalisieren lassen, z.B. Agenda-Setting oder Nachrichtenfaktoren (siehe Klein 1996). Bewertungsanalysen erfordern den Kontext von Suchbegriffen, und es gibt keine Software, die aufgrund semantischer Kriterien eine solche Analyse durchführen könnte. Ein weiterer Nachteil computerunterstützter Inhaltsanalysen liegt

[3] außer Hardware- und Softwarefehler

Abb. 1.3: Probleme der Arten von Inhaltsanalyse

konventionell	computerunterstützt
alle Arten von Inhaltsanalysen möglich	nur bestimmte Arten von Inhaltsanalysen möglich
Kategoriensystem für eine Textsorte gültig	Kategoriensystem nur für einen Materialkorpus gültig
geringe Transparenz des Kategoriensystems durch Beispiele und Kodierregeln	totale Transparenz des Kategoriensystems durch vollständige Aufzählung der Suchbegriffe
Kategorisierung und Kodierung weitgehend zusammenhängend	deutliche Trennung von Kategorisierung und Kodierung
Ermessensfreiraum beim Kodieren, dadurch Probleme durch Inter- und Intrakoderreliablität	bei automatischer Kodierung 100 Prozent Reliabilität[3]
Ambiguität und Negation werden von Kodierern erkannt	Ambiguität und Negation müssen aufwändig behandelt werden
kaum Transfer der Erfahrung, des methodischen *Wie*	akkumulative Forschung durch Publikation der Kategoriensysteme

darin, dass ein Kategoriensystem an dem Text entwickelt wurde, für dessen Analyse es eingesetzt wird. Andere Texte erfordern eine Erweiterung des Kategoriensystems. Diese Nachteile hat eine konventionelle Inhaltsanalyse nicht, dafür sind deren Kategoriensysteme kaum transparent, weil sie nur durch abstrakte Definitionen und an einigen Beispielen erläutert werden. Durch eine vollständige Aufzählung der Suchbegriffe ist bei einer computerunterstützten Inhaltsanalyse eine totale Transparenz gegeben. Kategorisierung und Kodierung lassen sich bei der konventionellen Inhaltsanalyse kaum trennen, während bei der computerunterstützten Inhaltsanalyse die Trennung sehr deutlich ist.

Beim Kodieren haben die Kodierer einer konventionellen Inhaltsanalyse einen Ermessensspielraum, der auch durch ein gutes Kategoriensystem und eine umfangreiche Kodiererschulung nicht ganz verschwindet. Daraus ergeben sich Unterschiede, wenn mehrere Kodierer mit dem gleichen Kategoriensystem denselben Text kodieren. Diese Unterschiede werden mit Koeffizienten für Interkoderreliabilität gemessen. Weiterhin treten durch Lerneffekte auch Unterschiede auf, wenn die gleiche Person den gleichen Text mehrmals kodiert und zwischen den Kodierungen eine größere Zeitspanne liegt. Dieses Phänomen wird Intrakoderreliabilität genannt. Sowohl Interkoder- als auch Intrakoderreliabilität verursachen einen Messfehler, der nie ganz verschwindet und eine 100prozentige Reliabilität der Messung unmöglich macht. Dem gegenüber hat eine automatische computerunterstützte Inhaltsanalyse keine Reliabilitätsprobleme, Messfehler infolge fehlerhafter Hard- oder Software treten so gut wie nie auf. Zwei wichtige Probleme der computerunterstützten Inhaltsanalyse, die aufwändig behandelt werden müssen, tauchen in kon-

ventionellen Inhaltsanalysen gar nicht auf, weil sie durch die Sprachkompetenz der Kodierer abgefangen werden: mehrdeutige und negierte Suchbegriffe. Bei ihnen muß eine Nachkodierung oder eine interaktive Kodierung erfolgen. Letztere ist aber zum einen nur in TEXTQUEST implementiert, zum anderen hat sie den Nachteil, dass dieselben Reliabilitätsprobleme wie bei der konventionellen Inhaltsanalyse auftreten. Betroffen sind davon nur die interaktiv kodierten Suchbegriffe und nicht die automatisch kodierten Suchbegriffe.

Zusammenfassung:

Die drei Übersichten haben gezeigt, dass beide Formen der Inhaltsanalyse ihre Vor- und Nachteile haben. Sie zeigen aber auch, dass gerade bei der Analyse massenmedialer Kommunikationsinhalte die computerunterstützte Inhaltsanalyse nicht nur für die Analyse großer Textmengen hervorragend geeignet ist, sondern auch einige andere Vorteile hat:

- 100 Prozent Reliabilität
- schnelle Kodierung, auch mehrfach
- große Textmengen und komplexe Kategoriensysteme sind analysierbar
- Transparenz des methodischen *Wie* durch Trennung von Kategorisierung und Kodierung und durch vollständige Offenlegung der Suchbegriffe
- Fehler beim Übertragen der Kodierergebnisse auf das Kodierblatt und von dort in einen Computer werden ausgeschlossen

1.5 Textanalytische Verfahren

Unter dem Begriff Textanalyse werden unterschiedliche Verfahren subsummiert, die im folgenden Abschnitt vorgestellt und diskutiert werden. Das Ziel ist, die Verfahren voneinander abzugrenzen, damit je nach Forschungsinteresse das angemessene Verfahren ausgewählt wird.

1.5.1 Hermeneutische Textanalyse

Bei der hermeneutischen Textanlayse geht es um die gesamte Interpretation eines Textes. Dabei werden alle verfügbaren Information über den Autor, die Zeit, in der er lebte, den Einflüssen, denen er ausgesetzt war, und möglichst alle verfügbaren Informationen bei der Interpretation des Textes berücksichtigt. Literaturwissenschaftler arbeiten so bei der Analyse von Gedichten, Romanen und anderen fiktionalen Texten. Ihr Erkenntnisinteresse liegt darin, möglichst alle Informationen bei der Interpretation zu berücksichtigen, sowohl die im Texte enthaltenen Informationen (textimmanent) als auch außerhalb des Textes (texttranszendent)

1.5.2 Qualitative Datenanalyse (QDA)

Seit den 80er Jahren des 20. Jahrhunderts – begünstigt durch die Einführung von PCs (persönlichen Computern) – ist eine Entwicklung von Software für die qualitative Datenanalyse festzustellen, z.B. atlas.ti, MaxQDA, The Ethnograph, HyperResearch und QDA-Miner[4]. Sie sind auch für die Informationsgewinnung (Text Retrieval) geeignet und insbesondere beim Auffinden bestimmter Textstellen nach bestimmten Kriterien und deren Kombinationen sehr schnell. Die Selektionskriterien gehen teilweise erheblich über die Möglichkeiten der Definition von Suchbegriffen bei den Programmpaketen für die computerunterstützte Inhaltsanalyse hinaus. Sie ermöglichen eine Kodierung von Textstellen, ohne vorher ein Kategoriensystem explizit festlegen zu müssen. Die Codes werden mit einem Editor in den zu analysierenden Text eingefügt. Die Art der Kodierung von Textstellen unterscheidet sich erheblich von der Kodierung einer computerunterstützten Inhaltsanalyse.

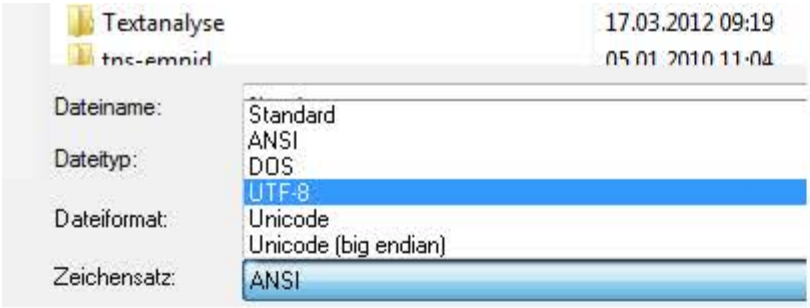
Einige Programme bieten ein Feature namens Autocoding an. Damit kann man ähnlich wie bei der computerunterstützten Inhaltsanalyse eine Kodierung von Suchbegriffen durchführen.

In diesem Buch soll es aber nicht um die QDA gehen. Eine Übersicht über die Unterschiede zwischen quantitativer und qualitativer Inhaltsanalyse findet sich bei Lamnek (Lamnek 2005: 272 und Atteslander 2010: 214-215).

[4] Eine ständig aktualisierte Übersicht befindet sich unter http://www.textanalysis.info

1.6 Exkurs: Texte, Zeichensätze und Encodierungen

Dieser Exkurs ist wichtig, wenn Texte in nicht-englischer Sprache analysiert werden sollen, die folgenden Ausführungen gelten nicht für Silbenschriften wie z.B. Chinesisch, Japanisch oder Koreanisch. Zur Bearbeitung der Textdateien nimmt man am besten einen Editor (z.B. Wordpad oder Textpad), beide können Textdateien mit UTF-8 Encodierung (für 8-bit UCS Transformation Format) speichern, wie die folgende Abbildung für Textpad zeigt:



Diese Form von Unicode ist heutzutage (2013) Standard und ist notwendig, weil viele Sprachen im Gegensatz zu Englisch besondere Zeichen wie Umlaute, Akzente oder Diakritika haben. Der Austausch von Texten zwischen verschiedenen Betriebssystemen stellt kein Problem mehr dar.

In den 80er und 90er Jahren des 20. Jahrhunderts gab es 7-bit Zeichensätze, die aber Umlaute oder Zeichen mit Diakritika nicht einheitlich darstellen konnten. Man musste dementsprechende Codepages auswählen, die Texte waren nicht austauschbar bzw. mussten konvertiert oder aufwändig bearbeitet werden, wenn sie diese Zeichen enthielten. Das änderte sich auch mit dem Aufkommen von MS-Windows Anfang der 90er Jahre nicht grundlegend, man sprach aber dann von ANSI-Datei (American National Standards Institute ANSI), einer 8-bit Zeichenkodierung.

Liegen die Texte in so einer Kodierung vor, müssen sie konvertiert werden. Die meisten Editoren können entsprechende Dateien einlesen und diese dann als UTF-8 Datei speichern.

Wichtig ist, dass alle Texte in derselben Encodierung vorliegen müssen. Je nach Software kann es Probleme geben, lediglich TextQuest kann sowohl UTF-8 als auch Latin-1 Encodierungen automatisch unterscheiden und verarbeiten. Ist die Encodierung unterschiedlich, findet man in einer Wörterliste z.B. Wörter, deren Umlaute nicht korrekt dargestellt werden.

1.7 Software für computerunterstützte Inhaltsanalysen

In diesem Buch werden die Möglichkeiten von TEXTQUEST vorgestellt und diskutiert. Es gibt weitere Programme, die TEXTQUEST ähnlich sind. Von Textpack gibt es eine Demoversion, die nur mit dem mitgelieferten Beispieltext funktionsfähig ist, nicht aber mit eigenen Texten. Eigene Texte können mit zeitlich begrenzenten Versionen von TEXTQUEST und WordStat ausprobiert werden. Unter http://www.textanalysis.info finden sich entsprechende Links.

Die Erstellung und die Aufbereitung der Texte erfolgt mit einem Editor oder einer Textverarbeitung außerhalb der Inhaltsanalyse-Software. Wenn die Texte als Datei vorliegen, müssen die Texte so aufbereitet werden, dass Texteinheiten und die Werte für externe Variablen definiert sind, dies geschieht meist durch Steuerzeichen.

SOZIALWISSENSCHAFTLICHE FORSCHUNGSMETHODEN

Ingwer Borg, Patrick J.F. Groenen, Patrick Mair:
Multidimensionale Skalierung

Band 1, ISBN 978-3-86618-438-1, München u. Mering 2010, 102 S., € 19.80

Carolin Strobl: **Das Rasch-Modell.**
Eine verständliche Einführung für Studium und Praxis

Band 2, ISBN 978-3-86618-695-8, München u. Mering,
2. erw. Aufl. 2012, 131 S., € 19.80

Jost Reinecke: **Wachstumsmodelle**

Band 3, ISBN 978-3-86618-692-7, München u. Mering 2012, 111 S., € 19.80

Jürgen H.P. Hoffmeyer-Zlotnik, Uwe Warner:
Soziodemographische Standards für Umfragen in Europa

Band 4, ISBN 978-3-86618-827-3, München u. Mering 2013, 124 S., € 19.80

Michael Stegmann, Julia Werner, Heiko Müller:
Sequenzmusteranalyse. Einführung in Theorie und Praxis

Band 5, ISBN 978-3-86618-829-7, München u. Mering 2013, 88 S., € 19.80

Dieses Buch liefert aus sozialwissenschaftlicher Perspektive einen Überblick über aktuelle Längsschnittdaten und schildert relevante Analyseverfahren. Der Schwerpunkt liegt auf der Durchführung der Sequenzmusteranalyse mit der Statistiksoftware TDA, STATA und R. Dabei wird die kostenlose und frei zugängliche Software R mit ihren vielfältigen Möglichkeiten der Längsschnittdatenanalyse in den Mittelpunkt gerückt. Die ausführlichen Beschreibungen der theoretischen Grundlagen der Sequenzmusteranalyse sowie die umfangreichen Vorgehenserläuterungen zur Deskription von Sequenzen, Optimal Matching und Clusteranalyse werden durch Beispiel-Skripte und zahlreiche Abbildungen ergänzt, sodass sie analog auf eigene Daten angewendet werden können.