Towards Extending Content Analysis (TECA) - Schlußbericht zu Arbeitspaket 1, Verschriftung

Geis, Alfons

Veröffentlichungsversion / Published Version
Abschlussbericht / final report

ZUMA-Technischer Bericht 98/18
ISSN 1437-4129

Towards Extending Content Analysis (TECA)
Schlußbericht zu Arbeitspaket 1
Verschriftung

Alfons Geis

ZUMA
Postfach 12 21 55
68072 Mannheim

Telefon: (06 21) 1246 - 225
Telefax: (06 21) 1246 - 100
E-Mail: geis@zuma-mannheim.de

ZUMA-Grundlagenforschungsprojekt TECA

TECA ist eine Pilotstudie, die das Methodenspektrum der Analyse von sozialwissenschaftlichen Texten erweitern soll. Dabei werden die Erfahrungen und Techniken aus der Linguistik, insbesondere der Computerlinguistik, sowie der angrenzenden Wissenschaften einbezogen und für die Sozialwissenschaften nutzbar gemacht.

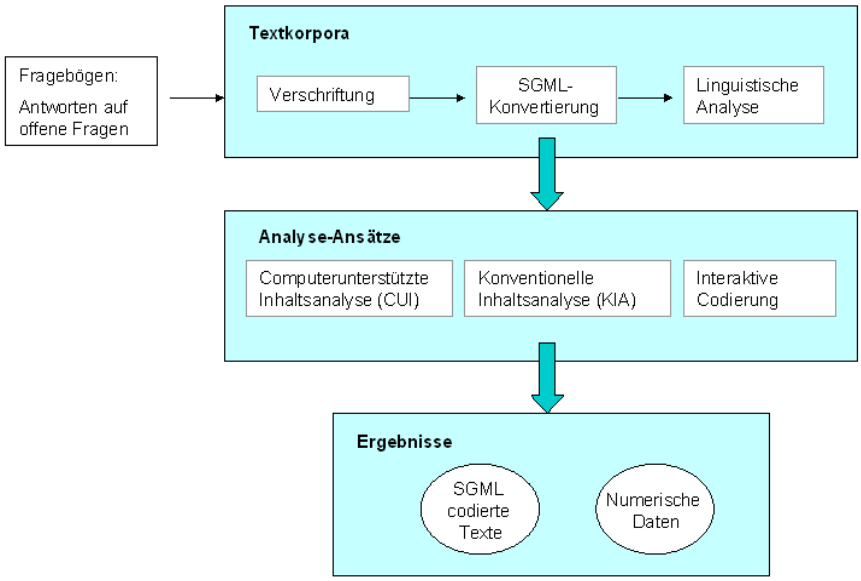
Eine Schwäche der Methode der *computerunterstützten Inhaltsanalyse* besteht darin, daß die Analyse überwiegend auf Einwort-Codierungen basiert, ohne den Kontext zu berücksichtigen. Diese fehlende Kontextsensitivität führt zu ungenauen Ergebnissen bzw. beschränkt die Anwendung der Methode der computerunterstützten Inhaltsanalyse auf bestimmte Themen und Textsorten. Ein wesentlicher Nachteil der *coderbasierten* (oder konventionellen) *Inhaltsanalyse* besteht u.a. in dem hohen Zeitaufwand und den aufwendigen Arbeitstechniken/-abläufen.

Lösungsmöglichkeiten sehen wir darin, daß einerseits zusätzliche linguistische Informationen bereitgestellt werden, die der computerunterstützten Inhaltsanalyse als „intelligente“ Komponente und der Codierkraft als leicht zugängliche Zusatzinformation dienen, andererseits versprechen wir uns in der Integration und gemeinsamen Anwendung der coderbasierten wie der computerunterstützten Inhaltsanalyse synergetische Effekte.

Als Untersuchungsmaterial wurde ein Texttyp gewählt, wie er in der Alltagspraxis von ZUMA am häufigsten vorkommt: Antworten auf offene Fragen aus zwei repräsentativen Stichproben. Sowohl kurze stichwortartige Antworten als auch längere argumentative sind vertreten. Für jede Stichprobe liegt ein Textkorpus vor. Die zwei Textkorpora werden mit Hilfe eines linguistischen Analysesystems (Parsers) bearbeitet. Dadurch wird jedes Textkorpus mit linguistischen Angaben angereichert, wie z.B. Wortstämmen, Wortartkategorien und syntaktischen Komponenten wie Verbphrasen, Nominalphrasen etc.

Im weiteren werden drei Analyse-Ansätze angewendet: die coderbasierte Inhaltsanalyse, die computerunterstützte Inhaltsanalyse und eine interaktive Codierung, die neben der Kombination der beiden ersten Verfahren auch das mit allen verfügbaren Informationen angereicherte Textkorpus nutzt. Die drei Ansätze werden miteinander verglichen und bewertet.

Ausgangspunkt für diese Analysen sind also Texte, die neben dem eigentlichen Text die Beschreibung der Textstruktur (inkl. projektspezifischer Daten), die maschinell und coderbasiert generierten Codes, die zugrunde liegenden Kategorienschemata und linguistischen (morpho-syntaktische) Merkmale enthalten, wie z.B. Wortkategorien, Verbphrasen, Nominalphrasen u.a. Damit werden „Multilevel-Analysen“ möglich, d.h. es können parallel verschiedene Informationsebenen ausgewertet werden. Außerdem versprechen wir uns davon eine Erweiterung des bisherigen Kategorisierungsspektrums sowie die Überprüfbarkeit und Präzisierung der bisherigen Codierung. In der folgenden Grafik wird der Analyseablauf für das Projekt TECA skizziert:



Die Anreicherung des Textes mit umfassender Information und die Verwendung für die unterschiedlichen Ansätze erfordern ein entsprechendes Textformat, das es erlaubt, neben dem eigentlichen Text alle vorhandenen und generierten Informationen verfügbar zu haben und wahlweise abrufen zu können. Das Textformat muß den Ansprüchen eines Standards genügen; nur so ist es über das derzeitige Projekt hinaus verwendbar. Seit einiger Zeit zeichnet sich eine Standardisierung auf dem Gebiet der Textdokumentation und -archivierung ab, nämlich SGML (Standard Generalized Markup Language). SGML ist plattform- und maschinenunabhängig, frei verfügbar (nicht-proprietär) und bietet eine genormte Vorgehensweise, eine Textstruktur zu spezifizieren: eine Document Type Definition (DTD) definiert mit Hilfe einer Reihe von Regeln die Struktur des Textes. Inhalt und Struktur bilden ein Ganzes, wobei die Strukturelemente und ihre Beziehungen zueinander durch die jeweiligen Dokumenttypen eindeutig beschrieben und definiert werden; eine solche DTD ist bereits für den in diesem Projekt verwendeten Texttyp erstellt und für die linguistischen Merkmale erweitert worden. Auch bezüglich der Texterfassung (Verschriftung) unmittelbar im SGML-Format liegen erste vielversprechende Erfahrungen vor.

Mit der Einführung eines Standards für Textformate wird gleichzeitig ein Hindernis beseitigt, indem die Austauschbarkeit von Texten zwischen verschiedenen Programmen und die Weitergabe von Texten zwischen Wissenschaftlern erleichtert wird.

Bei dem Projekt TECA handelt es sich um eine Machbarkeitsstudie, in der geprüft wird, mit welchem Aufwand welche Ergebnisse erzielt werden können und wie die verschiedenen Analyse-Ansätze zusammenwirken können (Abfolge, Unterstützung, Ergebnis). Die derzeit noch unbeantwortete Frage, inwieweit sich die gewonnenen Erfahrungen auch auf andere Texttypen (z.B. Leitfadengespräche oder Medientexte) übertragen lassen, wird am Ende zu diskutieren sein.

Für weitere Informationen beziehungsweise Anregungen wenden Sie sich an:
Dr. Melina Alexa (alexa@zuma-mannheim.de) und
Alfons Geis (geis@zuma-mannheim.de)

Inhaltsverzeichnis

1 Grundsätzliche Bedeutung der Verschriftung[1]

Grundvoraussetzung für eine Bearbeitung von Texten mit dem Computer ist die sogenannte Maschinenlesbarkeit, was bedeutet, daß die Texte als Datei vorliegen müssen. Viele Texte werden heute unmittelbar mit dem PC erstellt und als Datei weitergegeben. Ist dies nicht der Fall, so bleibt nur die Möglichkeit, die Texte einzuscannen oder abzuschreiben. Letzteres trifft vor allem dann zu, wenn handschriftliche Texte vorliegen. Wenngleich CATI (**C**omputer-**A**ided **T**elefon **I**nterview) oder CAPI (**C**omputer-**A**ided **P**ersonal **I**nterview) mit Laptop immer stärkere Verbreitung finden, werden in den Sozialwissenschaften - nicht zuletzt aufgrund des schwierigen Untersuchungsgegenstandes - noch sehr viele Interviews mit "Papier und Bleistift" durchgeführt, wobei die Anworten in einem persönlichen Gespräch notiert werden.

Betrachtet man den langen Weg der Informationsvermittlung vom Forscher über das Interview bis zu den Untersuchungsergebnissen, so ist die Anfälligkeit für unbeabsichtigte Veränderungen nicht zu übersehen, die in folgenden Fragen umschrieben werden können:

- Wird die Frage richtig formuliert; trifft sie das, was man wissen will; ist sie valide?
- Wird die Formulierung der Frage verändert, wenn sie der Interviewer stellt; erfährt der Sinn der Frage durch Auftreten, Gestik, Betonung o.a. des Interviewers eine Verfälschung?
- Wird die Frage akustisch und inhaltlich-intellektuell vom Befragten richtig verstanden?
- Kann sich der Befragte angemessen und unmißverständlich äußern?
- Wird die Antwort vom Interviewer richtig verstanden?
- Was ist von der Befragtenreaktion/-antwort festzuhalten? Schreibt der Interviewer die Antwort wortgenau auf oder wird der Sinn verändert?
- Sind die Notizen bei der Verschriftung unmißverständlich und eindeutig lesbar?
- Was von dem, was im Fragebogen steht, soll erfaßt werden? Wird der Text richtig und entsprechend den Regeln abgeschrieben?
- Was wird durch die Technik der Datenspeicherung verändert? Wie werden eine große, kleine Schrift, Betonungen, grafische Zeichen zur Strukturierung der Antwort u.v.a. behandelt?
- Wie verändern sich die Daten in der Auswertung mit den Analyseprogrammen durch Aggregation, Indexbildung, Mitteln u.a.?
- Welche Schlußfolgerungen werden aus den Daten gezogen? Ist die Interpretation zulässig und nachvollziehbar?

[1] Während noch vor einigen Jahren der Begriff "Verschriftung" für das Abschreiben schriftlicher oder gedruckter Vorlagen in einer maschnienlesbaren Form verwendet wurde und "Transkription" fast ausschließlich für die Übertragung von Tonaufnahmen in die Schriftform, wird heute "Transkription" immer mehr im Sinne einer Computererfassung jeder Art von Vorlage benutzt. Wir behalten die "Transkription" überwiegend der Erfassung von Audiodokumenten vor, wenngleich wir auch von "Transkribentinnen" sprechen, wenn sie schriftliche Textvorlagen erfassen (um das Wort Verschrifterinnen o.ä. unschöne Bezeichnungen zu vermeiden).

Ein Hauptkriterium der Wissenschaftlichkeit ist neben der Relevanz der Forschungsfrage und der Systematik die Nachvollziehbarkeit (Intersubjektivität). Diese erfordert eine genaue und dokumentierte Kontrolle jeder Veränderung.

Hinweise zur Erfassung schriftlicher Dokumente geben Züll, Mohler, Geis (1991, S. 37 - 48). Für mündliche Äußerungen, die deutlich komplexer sind, finden sich bei Mergenthaler (1992) umfangreiche Ausführungen. Der Forscher kann in den seltensten Fällen in allen oben genannten Phasen Einfluß und Kontrolle ausüben. Im vorliegenden Projekt TECA beginnt die Einflußmöglichkeit bei der Verschriftung der Antworten. Im folgenden sollen Texterfassung und -aufbereitung für die Computerbearbeitung dokumentiert werden.

2 Beschreibung der Texte

Für das Projekt TECA liegen zwei Textkorpora vor. Der erste Text stammt aus der Nachwahl-Studie von 1994. Es sind Antworten auf eine offene Frage; sie lautet für Ostdeutschland "Wenn Sie an die ehemalige DDR zurückdenken, gibt es Dinge, auf die Sie stolz sind?" und für Westdeutschland "Wenn Sie an die ehemalige DDR zurückdenken, gibt es Dinge, auf die die Menschen dort stolz sein können?". Die Intervieweranweisung, den Text genau zu notieren, wurde offensichtlich recht unterschiedlich interpretiert: die so verstandene "Genauigkeit" reicht von kurzen Stichworten bis zu ausführlichen Darstellungen in vollständigen Sätzen. Offen bleibt z.B., ob die deutlich knapperen Antwortnotizen in den westdeutschen Interviews auf die Befragten oder die Interviewer zurückzuführen sind. Auch das richtige Verstehen der Frage war offensichtlich nicht immer gewährleistet, wenn z.B. von persönlichen Erfolgen anstelle der gefragten Gemeinschaftsleistungen gesprochen wurde.

Verschriftet wurden die Antworten von dem Umfrageinstitut GFM-GETAS/WBA, die Verschriftungsregeln sind allerdings nicht mehr verfügbar. Das Dateiformat ist ASCII mit fester Datenstruktur, wie es mit der TEXTPACK-Funktion EXPORT (Mohler, Züll 1998) erstellt wird.

FB-ID Antworttexte

```
0009 Sozialleistungen
0010 Für alle Arbeit , soziale Dinge .
0014 Auf Familienbetreuung .
0016 Umwechslung von Ost- in Westmark , PDS ist steinreich .
0018 Kindergartensystem , keine Arbeitslosigkeit .
0019 Auf eigene Leistungen , das privat geschaffene .
0020 Auf die Organisation für berufstätige Mütter ( Kindertagesstätten )  ; geringe
     Arbeitslosenquote , geringe Kriminalität .
0022 Daß sie durch ihren Fleiß zu den stabilsten Faktoren im ehemaligen Ostblock zählen
     konnten .
0031 Soziale Einrichtungen , Familien- /Kinderbetreuung , Nachbarschaftshilfe
0032 In welcher Zeit sie diesen Status bekommen haben .
0038 Kinderhorte
0047 Daß sie aus eigener Kraft die Wende mitentschieden haben , die Schwierigkeiten
     gemeistert haben , die der Unterdrückung des SEDStaates zuzuschreiben war .
0050 Auf die Disziplin , die Kinderbetreuung ; auf die Frauen , die so aktiv waren .
```

Die Fragebogen-Nummern (IDs) 3000 bis 4133 beziehen sich auf die Interviews im Osten und IDs 1 bis 2999 auf den Westen. Es handelt sich um insgesamt 2046 Befragte (1046 Ost / 1000 West), von denen 1082 Fälle (794 Ost / 288 West) mit Antworten vorliegen; der Rest wurde ausgefiltert, weil auf die Vorfrage, ob es etwas gäbe, worauf sie stolz sein könnten, mit einem Nein geantwortet wurde.

Der zweite Text ist aus einem Forschungsprojekt von Petra Bauer-Kaase, das sie zusammen mit dem Wissenschaftszentrum Berlin für Sozialforschung (WZB) durchgeführt hat. Es trägt den Titel "Das Links-Rechts-Schema: Eine Analyse der inhaltlichen Bedeutung der Begriffe links und rechts". Durchführung der Interviews und Feldkontrolle wurden von GFM-GETAS/WBA vorgenommen[2]. Die Texte wurden uns freundlicherweise für unsere methodischen Fragestellungen zur Verfügung gestellt.

Die offene Frage lautete: "Können Sie mir bitte nun noch sagen, was Sie persönlich unter den Begriffen LINKS und RECHTS verstehen, wenn es um Politik geht?" Als Nachfragen - um die Antwortbereitschaft zu fördern - waren vorgesehen: "Und wäre sonst noch etwas dazu zu sagen?" bzw. "Und noch etwas, fällt Ihnen dazu noch etwas ein?" Unter der Rubrik "LINKS bedeutet" und "RECHTS bedeutet" waren von den Interviewern die Antworten der Befragten handschriftlich notiert worden. Die entsprechenden Fragebogenseiten lagen im Original vor.

Die Art der Notation bei der Verschriftung und technische Festlegungen erklären sich vor allem durch das für TEXTPACK (Mohler, Züll 1995) erforderliche Textformat - mit dem Programm-Paket TEXTPACK für computerunterstützte Textanalyse sollten die Texte weiterbearbeitet werden.

3 Verschriftungsregeln

Nach einer kurzen mündlichen Einweisung wurde den Transkribentinnen folgendes Regelwerk in schriftlicher Form zur Verfügung gestellt:

- Abzuschreiben sind die Antworttexte zur Frage 25 der GFM-GETAS-Studie U 6829/97/KA: "Können Sie mir nun noch sagen, was Sie persönlich unter den Begriffen LINKS und RECHTS verstehen, wenn es um Politik geht?"
- Die Texte sind in der Reihenfolge der Fragebogen-Nummern zu erfassen. Die Textvorlagen sind also vorher zu sortieren.
- Es werden zwei Identifikatoren vergeben: die Fragebogen-Identifikation (FB-ID) und die Kennung für die Frage:
  Die Fragebogen-Identifikation (FB-ID) wird mit einem "$" eingeleitet, gefolgt von der Ziffer, ohne daß ein Leerzeichen (Blank) eingefügt wird; die FB-ID findet sich auf jedem Blatt links oben, handschriftlich mit Filzstift. Führende Nullen brauchen nicht geschrieben zu werden.
  Die Frage-Nummer hat das Präfix "%", das ohne Blank vor die Frage-Nummer geschrieben wird, und zwar wird für die Antwort auf die Frage nach LINKS die Ziffer 1 und für die nach RECHTS die Ziffer 2 verwendet. Die Identifikationen "%1" und "%2" müssen für jeden Fragebogen vorliegen, auch wenn keine Angabe gemacht wurde: im letzteren Fall sind die entsprechenden Zeichen für fehlende Werte einzusetzen - vgl. unten.

[2] Nähere Projektinformationen sind dem Technischen Bericht 98/19, Toward Extending Content Analysis (TECA), Schlußbericht zu Arbeitspaket 2, Coderbasierte Inhaltsanalyse; Petra Bauer-Kaase, Alfons Geis zu entnehmen.

Für Anmerkungen zur Verschriftung steht die Frage-Nummer 3 zur Verfügung: Hier können alle Besonderheiten, Fragen und Anmerkungen, auch persönliche Kommentare zu dem entsprechenden Fragebogen festgehalten werden. Die Identifikation "%3" muß nur geschrieben werden, wenn auch eine Anmerkung gemacht wird.

Alle Identifikatoren stehen jeweils auf einer eigenen Zeile.
Beispiel:

$361	- Fragebogennummer
%1	- Kennung für Frage nach LINKS
offen für Neues, nicht so festgefahren	- Antwort des Befragten zu LINKS
%2	- Kennung für Frage nach RECHTS
konservativ, starr, rassistisch	- Antwort des Befragten zu RECHTS
%3	- Kennung für Kommentar
"konservativ" ist unterstrichen	- Anmerkung der Transkribentin

- Der Text wird unverändert abgeschrieben, ggf. fortlaufend über mehrere Zeilen hinweg. Was sich unter der Nachfrage "Und wäre sonst noch etwas dazu zu sagen? / Und noch etwas, fällt Ihnen dazu noch etwas ein?" findet, wird nicht getrennt erfaßt.
- Wird statt einer Anwort auf eine andere Stelle des Fragebogens verwiesen - "siehe oben", "vgl. 'Links' " o.ä. - so wird der Text, auf den verwiesen wurde, übernommen.
- Die Worte zu Beginn des Textes werden so geschrieben, als stünden sie mitten im Satz, können also auch klein beginnen.
- Aufzählungspunkte einer Liste oder eindeutig neue Gedanken/Angaben sind ggf. durch Komma oder ein sonstiges angemessenes Satzzeichen zu trennen.
- Unübliche, aber bekannte Abkürzungen werden ausgeschrieben. Sehr gebräuchliche Abkürzungen können übernommen werden, allerdings ohne Punkte. Die Abkürzungen müssen dann einheitlich geschrieben werden: zT, evtl, bzw, zB, usw, ua, uä, oä, uU, dh ... (Liste ist ggf. zu erweitern).
- Es wird lediglich die Rechtschreibung korrigiert, nicht aber Wortstellung, Satzbau oder Grammatik.
- Steht gar kein Text - leeres Feld, nur Strich o.ä. anstelle einer Anwort - im Fragebogen, werden die Fragebogen-ID, die Frage-Kennung und als Text "KA" geschrieben, ansonsten werden auch die Verweigerungen so geschrieben, wie vorgefunden; also z.B. "weiß nicht", "kann ich nicht sagen", "geht Sie nichts an" usw. Wenn sich bei der Nachfrage "Und wäre sonst noch etwas dazu zu sagen?" keine Angaben finden, zählt dies nicht als "fehlende Angabe".
- Hat der Interviewer ganz offensichtlich den Text in das falsche Feld geschrieben, z.B. rechts und links verwechselt (wenn z.B. bei rechts "linksradikal" und für links "die Nazis" steht), soll korrigiert werden; aber nur, wenn keinerlei Zweifel bestehen. Besteht der Verdacht einer Verwechslung, ohne daß man sich ganz sicher ist, so sollte dies unter der Identifikation "%3" notiert werden.

- Kann ein Wort nicht gelesen werden, so ist zunächst eine weitere Person zu fragen; kann das Wort auch dann nicht gelesen werden, so wird anstelle des Wortes "xyz" geschrieben.
- Um Doppeleingaben oder Mißverständnisse, besonders bei Verschriftung durch mehrere Personen, zu vermeiden, werden verschriftete Blätter links oben mit einem Handzeichen versehen.
- In die Arbeitsliste ist einzutragen, wer welche IDs zur Verschriftung wann übernommen hat; ebenso ist der Dateinamen zu vermerken, unter dem die Texte abgespeichert werden.
- Alle 30 bis 60 Minuten ist eine Datensicherung vorzunehmen.
- Der Dateiname ist "R-L-TXT1.DOC" für die Datei der ersten verschriftenden Person, "R-L-TXT2.DOC" für die zweite usw.
- In dem Text sollen keine Formatierungen vorgenommen werden; sie gehen im Zuge der weiteren Textaufbereitung für die computerunterstützte Textanalyse ohnehin verloren.

An der Verschriftung waren vier Personen in sehr unterschiedlichem Umfang beteiligt; deren Vorkenntnisse in bezug auf Erfassung von sozialwissenschaftlichen Texten von jahrzehntelanger Erfahrung bis absolutem Neuanfang reichten. So zeigte sich, daß vieles, was erfahrenen Kräften selbstverständlich war oder in Anlehnung an bisherige Verfahren entschieden werden konnte, in den Verschriftungsregeln nicht oder nicht eindeutig genug festgelegt war. Zu den offenen Fragen - und wie sie beantwortet wurden - zählten u.a. folgende:

- Wie werden graphische Zeichen wiedergegeben, die keine Buchstaben sind, z.B. das Gleichheitszeichen "=" oder Pfeile? -
  Je nach Kontext wurde "=" mit "gleich" bzw. "ist" oder auch nur durch ein Komma wiedergegeben.
  Ein Pfeil wurde (ebenso wie ein Bindestrich als Aufzählungszeichen) durch ein Komma wiedergegeben, im Sinne einer weiterführenden Aufzählung.
- Wie konsequent sollen Texte von den Stellen übernommen und wiederholt werden, auf die durch "s.o.", "vgl. links" u.ä. verwiesen wurde? Mit welcher Frage-Identifikation sollen pauschale Äußerungen versehen werden, die sich auf beide Fragen beziehen? -
  Während der Verschriftung wurden die Texte so geschrieben, wie sie vorgefunden wurden, die Fälle jedoch durch eine Bemerkung unter "%3" gekennzeichnet. Später wurden die Texte in Absprache mit der Vorgehensweise bei der manuellen Codierung dann wiederholt, wenn bei der Codierung ein zweites Mal darauf zugegriffen wurde.
- Wann darf die Zuordnung eines Textes zur Rubrik rechts oder links im Fragebogen beim Verschriften im Sinne einer Korrektur vertauscht werden, weil es sich offensichtlich um einen Irrtum des Interviewers beim Notieren handelt? -
  Da es sich zeigte, daß Befragte auch sehr widersinnige Angaben machen können, wurde die Plazierung unverändert gelassen, bis auf eine Ausnahme, wo der Antworttext so vollständig und eindeutig war, daß von einem Versehen des Interviewers gesprochen werden konnte.

- Wie sind Anführungszeichen und Hochkommata zu behandeln? -
  Teilweise waren ganze Antworten als Zitat des Befragten in Anführungszeichen gesetzt worden. Anführungszeichen wurden nur dann übernommen, wenn es sich um ein Zitat Dritter handelte oder der Begriff als ironisiert oder "sogenannt" gekennzeichnet werden sollte.

4 Textbereinigung und -aufbereitung

Nach Abschluß der Verschriftung wurde der gesamte Text mit dem Rechtschreibprüfprogramm von WORD Korrektur gelesen und als "Nur Text + Zeilenwechsel" abgespeichert.

Die mit "xyz" als nicht lesbar markierten Begriffe wurden von einer weiteren Person überprüft; auf diese Weise konnten von den ca. 30 unklaren Angaben zwei Drittel bereinigt werden. Während die Identifikation nicht lesbarer Stellen durch die xyz-Markierung kein Problem darstellte, war es schon schwieriger, die ungenau oder falsch gelesenen Stellen zu finden. Im Zuge der weiteren Textbearbeitung fiel nämlich auf, daß von den weniger in Verschriftung geübten und mit dem politikwissenschaftlichen Vokabular nicht so sehr vertrauten Personen die Handschriften teilweise ungenau interpretiert wurden. Deshalb wurde der gesamte Text mit den Originalnotizen im Fragebogen verglichen und ggf. korrigiert. Nur ein hohes Maß an "Schriftkenntnis" in Kombination mit dem Wissen um die relevanten feinen Unterschiede und Bedeutungen der Worte für die Codierung konnte hier Klarheit schaffen. Zu den bei nachlässiger Handschrift leicht verwechselbaren - von der Bedeutung aber manchmal recht unterschiedlichen - Begriffspaaren zählten u.a.: Mitte - Hitler, nur - mit, sozialistisch - sozial-liberal, ausländerfreundlich - ausländerfeindlich, selektiv - relativ, eben - aber, Nazisten - Rassisten, SDS - PDS, liberal - radikal, gilt - gibt, lange - bange u.a.m.

Desweiteren wurde die Zeichensetzung ergänzt, soweit es zum Verständnis der Texte notwendig schien, die Rechtschreibung überprüft, die Verwendung des Anführungszeichen bzw. Hochkommas und die Schreibweise der Abkürzungen vereinheitlicht.[3] Alle Anmerkungen der Transkribentinnen wurden durchgesehen, die darin angesprochenen Sachverhalte geklärt oder an anderer Stelle berücksichtigt, so daß alle Notizen schließlich gelöscht werden konnten und der reine Antworttext vorlag.

Weitere, formale Prüfungen folgten. Zunächst wurde mit Hilfe der Textaufbereitungsroutine SENTENCE von TEXTPACK die formale Textstruktur überprüft und ggf. in der Originaldatei korrigiert. Bei Erstellen einer Systemdatei für TEXTPACK wird u.a. angezeigt, wenn der Text nicht nach Identifikatoren sortiert ist. Dies ist bei sortierter Eingabe immer dann der Fall, wenn eine Fragebogen- oder Frage-Nummer falsch geschrieben wurde. Ebenso erfolgen Hinweise, wenn die Identifikation für Fragebogen oder Frage vergessen wurde, Text in der ID-Zeile steht

[3] Zur Bedeutung des Punktes in der weiteren Bearbeitung vergleiche den Technischen Bericht 99/01; Towards Extending Content Analysis (TECA), Schlußbericht zu Arbeitspaket 5, Morpho-syntaktische Analyse (Parsing) von deutschsprachigen Antworttexten; Beate Firzlaff und Michael Könyves-Tóth.

oder doppelte Fälle vorliegen. In etwa 1% aller Fragebögen war ein formaler Fehler dieser Art zu korrigieren, ein durchaus üblicher Umfang.
Der Text wurde über die Export-Routine von TEXTPACK in ein ASCII-Format mit fester Datenstruktur überführt. Das Präfix der Identifikatoren fiel weg, und die Textgliederungsmerkmale sind durch ihre Position in der Zeile definiert.

FB Frage Text

```
0001 1 SPD, sozial, human, legal
     2 CDU, egoistisch, wirtschaftlich im negativen Sinne für die
     2 Bevölkerung
0002 1 rot / grün
     2 christlich, altmodisch, auch radikal
0003 1 da kenne ich nur rot, sozial, fortschrittlich
     2 schwarz, bieder, altmodisch, christlich
0004 1 mir egal
     2 mir egal
0005 1 Vermittlung von sozial-liberal-emanzipatorischen Werten,
     1 Demokratisierung im kritischen Sinne der Gesellschaft
     2 christlich, konservativ, arbeitgeberfreundlich
```

Ein so strukturierter Text ist fast mit jedem beliebigen Anwendungsprogramm zu bearbeiten.
Nach Bereinigung von seiten des Projekts (ungültige oder unvollständige Interviews, doppelte Fragebogen-Nummern) standen 3022 Fälle mit Antworttexten zur Verfügung.

Zeitgleich mit der Verschriftung wurden die Texte konventionell codiert[4]. Die Codierdaten lagen zusammen mit den übrigen Umfragedaten (Interviewer-Nummer, Ort der Befragung, demographische Merkmale) als numerische Datei vor. Texte und zugehörige Codes wurden in einer einzigen Datei in der Weise zusammengefaßt, daß die Codes und zugehörige Texte je Frage unmittelbar nebeneinander standen.

```
FB   INT Codes links         Codes rechts       Frage-Nr./Text
-------------------------------------------------------------------------------------------------
0001138697141214121713571   7541951115216142    11SPD, sozial, human, legal
000113869                                        2CDU, egoistisch, wirtschaftlich im negativen Sinne für
000113869                                        2Bevölkerung
00021386971417151           156195115111        31rot / grün
000213869                                        2christlich, altmodisch, auch radikal
00031386971121414161        511553195111561     11da kenne ich nur rot, sozial, fortschrittlich
000313869                                        2schwarz, bieder, altmodisch, christlich
000413869                                        1mir egal
000413869                                        2mir egal
0005138692141155121113581   156145116511        31Vermittlung von sozial-liberal-emanzipatorischen Werten,
000513869                                        1Demokratisierung im kritischen Sinne der Gesellschaft
000513869                                        2christlich, konservativ, arbeitgeberfreundlich
0006180582141               5111                31sozialverträglich
```

Diese Datenform ermöglichte einerseits eine bequeme Überprüfung der manuellen Codierung, indem entweder nach Codes sortiert wurde oder bestimmte problematische Codes direkt ausgewählt wurden, andererseits stellte diese Datei die Ausgangsbasis für mehrschichtige Textannotation (Alexa 1999) dar, die darüber hinaus noch weitere Informationen enthalten soll, wie z.B. das Ergebnis der morpho-syntaktischen Analyse.[5]

[4] Dieser Vorgang ist in dem Technischen Bericht 98/19; Toward Extending Content Analysis (TECA), Schlußbericht zu Arbeitspaket 2, Coderbasierte Inhaltsanalyse; Petra Bauer-Kaase, Alfons Geis beschrieben.
[5] Vergleiche hierzu den Technischen Bericht 98/16; Towards Extending Content Analysis (TECA), Schlußbericht zu den Arbeitspaketen 4 und 6, Umsetzung in SGML-Format; Ingrid Schmidt, Melina Alexa

5 Zusammenfassung

Die vorliegenden Texte weisen einige Besonderheiten auf, die zu beachten sind, wenn an eine Textanalyse, insbesondere eine computerunterstützte, gedacht wird. Sie unterscheiden sich von "normalen" Texten (wie z.B. Zeitungsartikeln, Buchbeiträgen u.a.) dadurch, daß es sich um gesprochene Sprache handelt, die nicht die Gesetzmäßigkeiten der Schriftsprache aufweist; es gibt kaum vollständige Sätze, die Satzstrukturen sind inkonsistent, die Grammatik und der Wortgebrauch fehlerhaft. Es sind Spontanäußerungen aus Repräsentativumfragen, also von Personen aus allen Bevölkerungsschichten, die dazu noch meist laienhaft notiert wurden.

Den so entstandenen Texten fehlt die Regelhaftigkeit und Korrektheit, auf der viele linguistische Analyseinstrumente aufbauen. Nicht selten mit hoher Sprachkompetenz kaum verstehbar, bleibt damit der automatischen Weiterverarbeitung notwendigerweise ein Großteil der Information verschlossen.

Auch die beiden vorliegenden Texte sind nur bedingt miteinander vergleichbar:

- Die Fragestellungen fördern bei der Stolz-Frage eher stichwortartige Aufzählungen, während die Aufforderung einer Definition von LINKS und RECHTS auch Gelegenheit zu sehr ausschweifenden Erörterungen gibt. Und so sind auch die jeweiligen Texte im großen und ganzen charakterisiert.
- Wie die Links-Rechts-Texte in ein maschinenlesbares Format umgesetzt worden sind, war kontrollierbar, zu den Stolz-Texten liegen außer den Interviewer-Anweisungen keine Informationen vor. Die Erfahrung zeigt aber, daß "abschreiben, wie vorgefunden" sehr unterschiedlich umgesetzt werden kann, ohne daß gesagt werden könnte, was richtig oder falsch ist, solange keine expliziten Regeln vorliegen. So kann die Interviewernotiz sinngemäß abgeschrieben, nur die "wichtigen" Stichworte erfaßt oder buchstabengetreu wiedergegeben werden.
- Im Fall der Links-Rechts-Texte war die Überprüfung der Verschriftung etwas intensiver als gewöhnlich. Mit der manuellen Codierung war eine häufige und direkte Texteinsicht - und damit Überprüfung - verbunden.

Was bei sonst üblichen Verschriftungen dem Entscheidungsspielraum der Transkribentin überlassen werden kann, muß unmißverständlich und konsequent geregelt werden, wenn die Texte zu mehr als nur zu einer bisher angewandten computerunterstützten Inhaltsanalyse auf Wortbasis verwendet werden, wie z.B. bei der automatischen linguistischen Analyse (Parsing) oder zur Archivierung oder Nutzung im SGML-Format.

Literaturhinweise

Alexa, Melina und Schmidt, Ingrid (1999). Modell einer mehrschichtigen Textannotation für die computerunterstützte Textanalyse. In: Möhr, Wiebke und Schmidt, Ingrid (Hrsg.), SGML/XML - Anwendungen und Perspektiven, Heidelberg: Springer.

Mergenthaler, Erhard (1992). Die Transkription von Gesprächen. Ulm: Ulmer Textbank

Mohler, P.Ph., Züll, C. (1998). TEXTPACK PC, User Manual. Mannheim: ZUMA

Züll, C., Mohler, P. Ph., Geis, A. (1991). Computerunterstützte Inhaltsanalyse mit TEXTPACK PC. Stuttgart: Gustav Fischer