

## QUALSHOP - ZUMA-Workshop Datenmanagement bei qualitativen Erhebungsverfahren, März und Juni 1983: T. 1, Transkription, EDV-Aufbereitung und Analyse sprachlichen Materials

Veröffentlichungsversion / Published Version

Sammelwerk / collection

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

### Empfohlene Zitierung / Suggested Citation:

Zentrum für Umfragen, Methoden und Analysen -ZUMA-. (1984). *QUALSHOP - ZUMA-Workshop Datenmanagement bei qualitativen Erhebungsverfahren, März und Juni 1983: T. 1, Transkription, EDV-Aufbereitung und Analyse sprachlichen Materials* (ZUMA-Arbeitsbericht, 1984/01). Mannheim. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-70408>

### Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

### Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

# ZUMA

Zentrum für Umfragen, Methoden und Analysen e. V.

Postfach 5969

6800 Mannheim

B 2, 1

Telefon (06 21) 111111

---

## QUALSHOP

ZUMA-Workshop Datenmanagement  
bei qualitativen Erhebungsverfahren  
März und Juni 1983

### Teil 1

Transkription, EDV-Aufbereitung  
und Analyse sprachlichen Materials

ZUMA-Arbeitsbericht 84/01

Januar 1984

- Sammlung von Arbeitspapieren  
und Berichten -

---

Geschäftsführender  
Direktor

**Manfred Küchler**  
Vorsitzender  
Max Kaase  
Universität Mannheim

Mitglieder  
Hubert Feger,  
Universität Hamburg

Walter Jaide,  
Forschungsstelle für  
Jugendfragen, Hannover

M. Rainer Lepsius,  
Universität Heidelberg

Franz Urban Papp,  
Universität Kiel

Hansgerit Peiserl,  
Universität Konstanz

Erwin K. Scheuch,  
Universität zu Köln

Werner Tack,  
Universität Saarbrücken

Rudolf Wildenmann,  
Universität Mannheim und  
Europäisches Hochschulinstitut Florenz

Wolfgang Zapf  
Universität Mannheim

Rolf Ziegler,  
Universität Mur

Ab Juli 1983 sind die bisherigen ZUMA-Arbeitsberichte in zwei Reihen aufgeteilt:

Die ZUMA-Arbeitsberichte (neue Folge) haben eine hausinterne Begutachtung durchlaufen und werden vom Geschäftsführenden Direktor zusammen mit den übrigen wissenschaftlichen Leitern herausgegeben. Die Berichte dieser Reihe sind zur allgemeinen Weitergabe nach außen bestimmt.

Die ZUMA-Technischen Berichte dienen dem Zweck der hausinternen Kommunikation bzw. der Unterrichtung externer Kooperationspartner. Sie sind nicht zur allgemeinen Weitergabe bestimmt.

# Inhaltsverzeichnis

<u>Teil I</u>		<u>Seite</u>
Einleitung	M. Kuchler/P.Ph. Mohler	1
<u>1. Texte transkribieren oder abschreiben und für die EDV aufbereiten</u>		3
1.1 Die Transkription von Gesprächen Eine Zusammenstellung der Regeln	E. Mergenthaler	5
1.2 Textaufbereitung für die EDV- unterstützte Textanalyse in den Sozialwissenschaften	P.Ph. Mohler	27
1.3 Anleitung zum 'Direct Data Entry'	B. Degenhardt	45
<u>2. Textmanipulationen und Textanalysen</u>		
2.1 SAR-Sample And Retrieval	P.Ph. Mohler	53
2.2 TEXTPACK V- Kurzbeschreibung	P.Ph. Mohler/C. Züll	63
2.3 Möglichkeiten EDV-gestützter Strukturierung und Ordnung des qualitativen Interviews	J. Schupp	93
<u>3. Spezielle Anwendungen: Halboffene Vercodung als Mittel der Datenreduktion</u>		120
3.1 Projektbericht: Systematische Aufarbeitung des Archivs des Sigmund-Freud-Instituts in Frankfurt	P.Ph. Mohler	121

## Inhaltsverzeichnis

### Teil 2

		<u>Seite</u>
Einleitung	M. Küchler/P.Ph. Mohler	1
1. SIR/DBMS in der qualitativen Forschung	B. Degenhardt/ W. Degenhardt	5
2. TEXT BASE MANAGEMENT Systeme - Werkzeuge zur Archivierung und Analyse sprachlicher Daten	E. Mergenthaler	56
3. Planung und Entwicklung eines Programmsystems zur EDV-gestützten qualitativen Analyse von Verbaldaten	W. Paulus	65
4. Ungenauigkeit und Effizienz Die Informationsstruktur hi- storischen Quellenmaterials und ihre Bearbeitung mit dem datenbankorientierten Pro- grammsystem CLIO	M. Thaller	91

### Anhänge

1. Programme beider Workshops		124
2. Referenten und Teilnehmer Qualshop I		129
3. Referenten und Teilnehmer Qualshop II		135

Anhänge

Seite

1. Programme beider Workshops	141
2. Referenten und Teilnehmer Qualshop I	149
3. Referenten und Teilnehmer Qualshop II	155



## Einleitung

Die technischen Möglichkeiten der Bild- und Tonaufzeichnung erlauben mit verhältnismäßig einfachen Mitteln, große Mengen sprachlichen Materials aufzuzeichnen. Dagegen haben die angewandten Techniken der Texterschließung und der Textanalyse in der Regel nicht den entsprechend hohen Stand der Texterhebung erreicht. Papier, Schere und Zettelkasten als klassische Hilfsmittel der Texterschließung sind ebenso wie nicht-kodifizierte Verfahren der Textanalyse der Überfülle gesammelten sprachlichen Materials nicht mehr gewachsen. Zumindest für die Texterschließung stehen jetzt schon Computertechniken bereit, die wesentliche Erleichterungen bei der Verwaltung großer Textbestände bieten. Angefangen bei automatisch erzeugten Textvokabularen über die Indexerstellungen und Begriffssuche bis zu komplexen Suchprozessen, stehen derzeit ausgezeichnete Hilfsmittel zur Verfügung, die bei entsprechender Kodifizierung der Regeln auch als Hilfsmittel der Textanalyse eingesetzt werden können.

Die überaus starke Resonanz auf unsere Ankündigung, im Frühjahr 1983 bei ZUMA einen Workshop zum Thema "Datenmanagement bei qualitativen Erhebungsverfahren" zu veranstalten, bestätigte unsere Einschätzung, daß besonders in qualitativ ausgerichteten Forschungsprojekten ein erheblicher Bedarf an diesen neuen Techniken besteht.

Auf dem Workshop haben Forscher aus verschiedenen sozial- und geisteswissenschaftlichen Bereichen über ihre Erfahrungen mit computerunterstützten Techniken der Texterschließung und der Textanalyse berichtet.

Einen großen Teil der dort vorgelegten Berichte haben wir jetzt in zwei Bänden zusammengefaßt. Zusätzlich haben wir andere Arbeitspapiere, die uns als systematische Ergänzungen für gut erschienen, in die Sammlung mit aufgenommen.

Die Teilung in zwei Bände lag aus thematischen Gründen nahe. Im ersten Teil werden Probleme der Übertragung sprachlichen Materials in computergemäße Formen, einfache und elaborierte Texterschließungs- und Analyseverfahren behandelt.

Die dort vorgestellten Lösungen dürften für die Behandlung sogenannter "Einwegtexte" ausreichend sein. "Einwegtexte" meint, daß das sprachliche Material nur für einen bestimmten Forschungszweck und zur einmaligen Analyse aufbereitet werden soll.

Die Probleme, allgemeineren Zwecken dienliche Text-Daten-Basen bereitzustellen, werden dann im zweiten Band dieser Sammlung erörtert.

Wir haben die einzelnen Beiträge zu thematischen Gruppen zusammengefaßt, denen jeweils eine kurze Einleitung vorangestellt ist.

Ein Verzeichnis der Autoren und Teilnehmer beider Workshops, das wir jedem Teil beigeben, soll den direkten Kontakt zwischen interessierten Forschern erleichtern.

Manfred Küchler

Peter Ph. Mohler

## 1. Texte transkribieren oder abschreiben und für die EDV aufbereiten

Der erste Schritt einer systematischen Analyse gesprochener Sprache ist deren Überführung in eine wie auch immer notierte Schriftform. Die Art der Transkription legt spätere Analysemöglichkeiten weitgehend fest. Paraverbale Äußerungen wie z.B. "hm, äh" sind für Diskursanalysen ein wichtiges Merkmal, in rein thematisch orientierten Inhaltsanalysen sind sie eher schmückendes Beiwerk. Ähnliches gilt für die Aufzeichnung nicht verbaler Äußerungen und situativer Geräusche. Je feiner die Transkription durchgeführt wird, desto mehr Aufwand kostet dies. Hierbei ist insbesondere der hohe zeitliche Aufwand ein wesentliches Hindernis, große Mengen sprachlichen Materials zu transkribieren. Man kann als Faustregel angeben, daß eine rein auf verbale Äußerungen und Hochsprache beschränkte Transkription zwischen 7 und 11 Stunden Transkriptionszeit pro Stunde gesprochener Sprache erfordert.

Die Überführung von Transkriptionen in Computer (maschinenlesbare Texte) erfordert die Einhaltung weiterer Regeln, die dann auch für die Aufbereitung schon geschriebener Texte gültig sind.

Der folgende Beitrag E. Mergenthalers beschreibt die in Ulm festgelegten Transkriptionsregeln und die Schreibvorschriften für OCR (Klarsicht)-Belegleser.

Der anschließende Beitrag P. Mohlers erörtert Probleme der EDV-Aufbereitung von Texten wobei er sich auf einfachste Transkriptions- bzw. Abschreiberegeln beschränkt. B. Degenhardt hat eine Anweisung für die Schreibkräfte entwickelt, aus der implizit die Probleme der praktischen Umsetzung von Transkriptionsregeln hervorgehen.



Die Transkription von Gesprächen

Eine Zusammenstellung der  
Regeln

Erhard Mergenthaler

Arbeitspapier

Januar 1984

Abteilung Psychotherapie

SFB 129 - Projekt E2

Universität Ulm  
Am Hochsträß 8

D 7900 Ulm

## Inhaltsverzeichnis

## Einleitung

## Teil A Allgemeine Hinweise zur Transkription

1. Verbale Äußerungen
2. Paraverbale Äußerungen
3. Nicht verbale Äußerungen
4. Situationsgebundene Geräusche

## Teil B Kennzeichnung der Gesprächseinheiten

1. Hervorhebung
2. Namen
3. Zitate
4. Sprachwechsel
5. Satzzeichen
6. Gleichzeitigkeit
7. Wortabbrüche
8. Unverständliche Redeteile
9. Pausen
10. Zeitangaben
11. Interjektionen
12. Mehrdeutigkeit
13. Kommentar
14. Worttrennung
15. Diskontinuierliche Formen
16. Auslassungen
17. Umlaute und ß
18. Klein- und Großschreibung
19. Abkürzungen
20. Stottern
21. Zahlen, Brüche etc.
22. Wortschöpfungen
23. Fehlleistungen
24. Allgemeines

## Teil C Transkriptionsstandard

## Teil D Druckaufbereitung der Transkripte

## Teil E Hinweise zum Schreiben von Texten

1. Das Erstellen von Belegen
2. On-Line-Transkription am TELECOMP

## Teil F Übersicht zur Verwendung der Zeichen

## Literaturhinweise

## Einleitung

Die nachfolgende Zusammenstellung von Regeln für die Transkription von Gesprächen orientiert sich an drei Gesichtspunkten:

1. Leichte Lesbarkeit der erstellten Transkripte.
2. Möglichst geringer Aufwand bei der Transkription.
3. Computerverträglichkeit der Transkripte.

Alle drei Forderungen ließen sich nur durch Kompromisse erfüllen. Dies betrifft insbesondere die mögliche Verwendung der Transkripte für wissenschaftliche Untersuchungen. Eine Orientierung fand hierzu an den Fragestellungen statt, wie sie bei der Erforschung psychotherapeutischer Texte auftreten. Spezielle linguistische Probleme, etwa die Erforschung von Intonationsphänomenen, lassen sich hiermit nicht bearbeiten.

Unter forschungsökonomischen Gesichtspunkten war es notwendig, das Regelwerk so zu gestalten, daß Schreibkräfte bereits im ersten Durchgang qualitativ hochwertige Transkripte erstellen und auch gegebenenfalls den zweiten Arbeitsschritt des Korrekturhörens übernehmen können. Angesichts der knappen Ressourcen an wissenschaftlichem Personal kommt diesem Gesichtspunkt besondere Bedeutung zu.

Die Computerverträglichkeit schließlich sollte sicherstellen, daß die Weiterverarbeitung der Transkripte keine allzu großen Softwareprobleme aufwirft. Es wurde deshalb, um ein Beispiel zu nennen, auf die teilweise übliche Partiturschreibweise verzichtet. Dennoch ist es möglich, die erfaßten Phänomene im Nachhinein, etwa durch spezielle Druckprogramme, in partiturartige Druckbilder umzusetzen.

Da die ULMER TEXTBANK einen großen Anteil ihrer Texte von Institutionen außerhalb Ulms bezieht, die zumeist über keinen Computer verfügen, wurde das Verfahren zur Datenübertragung auf maschinenlesbaren Belegen weiterhin als fester Bestandteil in die Regeln mit einbezogen. Darüber hinaus besteht aber die Möglichkeit, die technologischen Fortschritte der jüngsten Zeit in Anspruch zu nehmen und über Mikrocomputer oder Textverarbeitungssysteme die Transkription on-line mit der entsprechenden Unterstützung, etwa zur Schreibfehlerkorrektur, durchzuführen. Diese Entwicklung wird zukünftig verstärkt verfolgt werden, indem die ULMER TEXTBANK mobile Texterfassungstationen einzusetzen plant.

## A Allgemeine Hinweise

In das Transkript werden alle auf der Tonaufzeichnung hörbaren Zeichen aufgenommen. Im einzelnen sind dies die verbalen, die paraverbalen und die nicht-verbalen sprachlichen Äußerungen der beteiligten Sprecher sowie die situationsgebundenen Geräusche.

### 1. Verbale Äußerungen

Hierunter fallen alle ganz oder teilweise ausgesprochenen Wörter. Sie werden in literarischer Umschrift wiedergegeben, die sich an der Schreibweise der Schriftsprache orientiert, jedoch nur die tatsächlich hörbaren Laute berücksichtigt. Dialekt wird in der entsprechenden hochsprachlichen Form transkribiert (siehe aber Regel B 4). Hier einige Beispiele:

-----  
is, hab, hab's, ne, gschafft  
-----

### 2. Paraverbale Äußerungen

Dazu zählen alle Laute oder Lautfolgen, die meist für sich allein, also in keinem Satzgefüge geäußert werden und als Pausenfüller, als Ausdruck des Zweifels, Bestätigens, der Unsicherheit, des Nachdenkens usw. dienen. Sie werden wie Sätze behandelt und ebenfalls in literarischer Umschrift festgehalten. Hier einige Beispiele:

-----  
hm, äh, ui, s, f  
-----

### 3. Nicht verbale Äußerungen

Hierzu gehören alle sonstigen geräuschvollen Sprecherhandlungen. Sie werden als Kommentar an der Stelle ihres Auftretens in das Transkript aufgenommen.

-----  
(hustet heftig), (lacht), (stoßhnt)  
-----

### 4. Situationsgebundene Geräusche

Sie entstehen durch die Umwelt und gehören mit zur Sprechsituation. Sie werden mit Hilfe von Kommentaren festgehalten. Hier einige Beispiele:

-----  
(Telefon klingelt), (Fluglaßrm)  
-----

## B Kennzeichnung der Gesprächseinheiten

Zur Kennzeichnung weiterer Gesprächsbesonderheiten werden Sonderzeichen eingeführt und in den Text eingefügt. Mit ihnen kann die Hervorhebung besonderer Redeteile, die Kennzeichnung von Zitaten sowie der Wechsel in der Sprechweise vermerkt werden. Eine weitere Gruppe von Zeichen, die Satzzeichen, werden verwendet um den Rhythmus der Rede zu markieren. Nachfolgend wird die Verwendung der Zeichen im einzelnen erläutert.

### 1. Hervorhebung und Dehnung

Wortformen, die vom Sprecher durch deutliche Betonung hervorgehoben sind (Emphase), werden durch ein nachgestelltes Ausrufezeichen (s.S.13 Sonderzeichen) gekennzeichnet. Dehnt der Sprecher eine Wortform auffällig, dann wird ein Doppelpunkt nachgestellt. Hier ein Textausschnitt mit je einer Dehnung und einer Betonung:

-----  
 also: wenn Sie das! tun  
 -----

### 2. Namen

Eigennamen, Städtenamen und Landschaftsbezeichnungen werden durch Ersatznamen (Pseudonyme) ersetzt und mit einem vorangestellten Stern gekennzeichnet. Innerhalb eines Textes sollten immer dieselben Pseudonyme verwendet werden.

-----  
 gestern meinte \*Paul, das\$ Herr \*von-Tor  
 -----

### 3. Zitate

Wörtliche Reden und Zitate werden zur Abgrenzung der vom Sprecher sonst hervorgebrachten Redeteile in Hochkommata eingeschlossen.

-----  
 da schrie er 'las\$ mich in Ruh' oder so was  
 -----

### 4. Sprachwechsel

Wechselt ein Sprecher für wenige Wörter seine Sprache von seiner gewohnten und für ihn üblichen (habituellen) in eine andere, dann werden diese in Anführungszeichen eingeschlossen und in literarischer Umschrift wiedergegeben. Mit einem nachgestellten Kommentar kann die Variante näher bezeichnet werden.

-----  
 dieses ewige "Gschnader" (schwaßbisch) haßngt mir  
 -----

## 5. Satzzeichen

Sie werden bei allen rhythmischen und syntaktischen Einschnitten des Redeverlaufs entsprechend dem Tonhöhenverlauf gesetzt. In Satzzeichen eingeschlossene Redeteile werden Kadenzen genannt. Folgende Zeichen finden Verwendung:

- ? steigende/hoch endende Kadenz
- = schwebende/in mittlerer Lage endende Kadenz
- , fallende/halbtief oder
- . auf dem Grundton endende Kadenz
- ; nicht zu Ende gesprochene Kadenz

Während die Bedeutung der Zeichen . und , damit weitgehend dem Gebrauch in der Schriftsprache entsprechen, wird für das Zeichen ? hier eine abweichende Gebrauchsweise vereinbart, indem es ausschließlich für die Stimmhebung verwendet wird. Dies stimmt nicht immer mit der von der Schriftsprache her gebräuchlichen Verwendung dieses Zeichens zur Markierung von Fragen überein.

Satzzeichen werden also überall dort gesetzt, wo aufgrund eines Innehaltens, einer Pause oder eines TonhöhenSprungs oder -umschwungs eine Kadenz erkannt wird.

Die Satzzeichen werden dem vorausgehenden Wort unmittelbar angehängt. Danach wird eine Leerstelle eingefügt. In dem nachfolgenden Textausschnitt fängt der Sprecher zunächst einen Satz mit "Sie sagten" an. Dann bricht er jedoch ab und beginnt einen neuen Satz mit "wissen Sie":

-----  
 nein= nein. oder doch? Sie sagten; wissen Sie,  
 -----

## 6. Gleichzeitigkeit

Reden zwei Sprecher gleichzeitig, dann wird dies im Transkript ebenfalls vermerkt. Da die Gleichzeitigkeit in der Schriftform insbesondere im Hinblick auf computerunterstützte Auswertungsmethoden nur hintereinander zum Ausdruck gebracht werden kann, wird zunächst beim ersten Sprecher der Beginn der Gleichzeitigkeit durch ein vorangestelltes Pluszeichen markiert und mit dessen Redebeitrag solange fortgefahren, bis die Gleichzeitigkeit beendet ist. Hier wird ein Sprecherwechsel notiert und der Redebeitrag des zweiten Sprechers bis zum Ende der Gleichzeitigkeit transkribiert. Der Abschluß erfolgt mit einem nachgestellten Pluszeichen. Behält der zweite Sprecher das Wort, so wird die Transkription ohne Sprecherwechsel weitergeführt. Im anderen Falle wird mit dem Redebeitrag des ersten

Sprechers nach der Gleichzeitigkeit fortgefahren. Diese Markierung erfolgt immer auf Wortgrenze, auch wenn die Gleichzeitigkeit nur auf Wortteile beschränkt ist. Im folgenden Beispiel wird "A/" und "P/" als Sprecherkennung verwendet:

-----  
P/ ich habe da eine +Venenerkrankung  
A/ ja= ja.+ das sagten Sie schon mal  
-----

## 7. Wortabbrüche

Ein nicht zu Ende gesprochenes Wort, sei es durch den Sprecher selbst oder durch das Dazwischenreden eines anderen Gesprächsteilnehmers verursacht, wird durch einen unmittelbar vor- oder nachgestellten Bindestrich gekennzeichnet. Hier ein Textausschnitt:

-----  
merkwaß- ich hab das Gefußhl, -fußhl  
-----

## 8. Unverständliche Redeteile

Für jedes unverständliche Wort wird ein Schrägstrich in das Transkript aufgenommen. Sind mehrere Wörter unverständlich, dann werden so viele Schrägstriche - jeweils durch eine Leerstelle getrennt - aufgenommen, wie Wörter ausgemacht werden können. Hier ein Textausschnitt:

-----  
da habe ich / / / wie es / sein  
-----

Anstatt der Markierung der unverständlichen Redeteile kann der Transkribent den von ihm vermuteten Wortlaut (Emendation) in ein doppeltes Klammernpaar einfügen. Beispiel:

-----  
gestern ((vormittag)) habe ich  
-----

## 9. Pausen

Sprechpausen werden mit einem oder mehreren aufeinanderfolgenden Bindestrichen gekennzeichnet. Die Anzahl der Bindestriche ist abhängig von der Dauer der Pause und kann aus folgender Tabelle entnommen werden.

Dauer der Pause	Anzahl Bindestriche	Zuwachs/Strich
2 sec	1 -	2 sec
5 sec	2 --	3 sec
10 sec	3 ---	5 sec
15 sec	4 ----	5 sec
ab hier Stoppuhr verwenden		
30 sec	5 -----	15 sec
1 min	6 -----	30 sec
2 min	7 -----	1 min
5 min	8 -----	3 min
10 min	9 -----	5 min
15 min	10 -----	5 min
30 min	11 -----	15 min
mehr	12 -----	

Aus der Tabelle geht hervor, daß Pausen unter 2 Sekunden nicht vermerkt werden. Für ein praktisches Vorgehen empfiehlt es sich, zunächst bis zu einer Dauer von 15 Sekunden den Zuwachs durch Zählen, wie z. B. "21,22,23" für 3 Sekunden, abzuschätzen und danach den nächsten Bindestrich zu setzen. Dauert eine Pause länger, so empfiehlt sich der Gebrauch einer Stoppuhr.

Allgemein ist zu beachten, daß der nächste Bindestrich erst gesetzt wird, wenn die nächste Grenze erreicht ist, das Gespräch also während der vollen Dauer des jeweiligen Zuwachses nicht fortgesetzt wurde. Als Beispiel folgt ein Textauschnitt:

-----  
 ich weisS nicht, --- heute  
 -----

Pausen werden grundsätzlich bei dem Gesprächsteilnehmer notiert, der zuletzt gesprochen hat. Dies gilt insbesondere dann, wenn nach der Pause ein anderer Sprecher fortfährt.

#### 10. Zeitangaben

Um die seit Beginn eines Gesprächs verfllossene Zeit zu kennzeichnen, wird die Form des Kommentars gewählt. Als Kennzeichnung wird nach der öffnenden Klammer ein Pluszeichen eingefügt. Damit läßt sich die relative Zeit zum Gesprächsbeginn notieren. Minuten werden durch ein der Zahl nachgestelltes Hochkomma, Sekunden durch ein Anführungszeichen gekennzeichnet. Hier ein Beispiel:

-----  
 (+3'10")  
 -----

#### 11. Interjektionen

In der gesprochenen Sprache bilden Interjektionen meist selbst-

ständige Kadenzen. In Verbindung mit den Satzzeichen können die jeweiligen Bedeutungsunterschiede bei deren Gebrauch festgehalten werden. Als Beispiel dienen die in folgender Tabelle aufgeführten Varianten für das Partikel "hm".

Verwendung	Bedeutungen
hm hmhm	Bestätigung
hm?	Frage
hm= hmhm=	Zögern
hm, hmhm,	Verwunderung
hm. hmhm.	Ratlosigkeit
hmhm-	Verneinung

## 12. Mehrdeutigkeit

Im Hinblick auf computerunterstützte Auswertungen kann es nützlich sein, mehrdeutige Wortformen durch entsprechende Markierungen eindeutig werden zu lassen. Hierzu können einer beliebigen Wortform hinter einem nachgestellten Schrägstrich beliebige Zahlen oder weitere Wortformen angehängt werden. Mit den Zahlen werden Bedeutungsvarianten einer Wortform gekennzeichnet. Mit den nachgestellten Wortformen können beim Gebrauch der Pronomina die Referenten genannt werden. Als Hilfsmittel kann das Basiswörterbuch verwendet werden. Beispiele:

-----  
Pank/2, wir/Gruppe, dort/Ulm  
-----

## 13. Kommentar

Neben den bis hierher vereinbarten festen Kommentaren können beliebige freie Kommentare verwendet werden. Sie werden in runde Klammern eingeschlossen und können alle Zeichenfolgen enthalten, die nicht durch die oben vorgegebenen Anwendungsfälle verbraucht sind.

## 14. Worttrennung am Zeilenende

Zur Trennung einer Wortform am Zeilenende wird dem ersten Wortteil ein Prozentzeichen nachgestellt und in der nächsten Zeile mit dem zweiten Wortteil fortgefahren. Die Worttrennung braucht nicht den üblichen Regeln der Silbentrennung zu entsprechen. Als Beispiel hier ein Textausschnitt:

-----  
die Zeiten fu\$ den Lad\$  
enschlus\$ haben sich nun  
-----

## 15. Diskontinuierliche Formen

Besitzen zwei Wortformen denselben Wortstamm, jedoch unter-

schiedliche Vor- bzw. Nachsilben, so wird in der gesprochenen Sprache häufig bei dem einen Wort der Wortstamm weggelassen, wenn beide Wortformen mit einer Konjunktion verknüpft werden. In diesem Fall wird der weggelassene Wortstamm durch einen angehängten Gedankenstrich angedeutet. Diese Regelung gilt auch für weggelassene Vor- oder Nachsilben. Hier zwei Beispiele:

-----  
An- und Abflug, Hauptfrage und -antwort  
-----

#### 16. Auslassungen

Das in der Schriftsprache gebräuchliche Auslassungszeichen wird bei Transkripten gesprochener Sprache entsprechend der literarischen Umschrift nur dann verwendet, wenn zwei verschiedene Wortformen zu einem Laut zusammengezogen sind. Es wird also nur verwendet, wenn es zwischen zwei Buchstaben eingeschlossen ist. Nach einem unmittelbar vorauslaufenden Leerzeichen muß es entfallen. Beispiele:

-----  
hat's, so'n aber so ne, s geht  
-----

#### 17. Umlaute und ß

Sofern auf der verwendeten Tastatur nicht vorhanden, gilt folgende Regelung: Das sogenannte "scharfe s" oder "ß" wird wie s\$ verschriftet. Umlaute wie a\$, o\$, und u\$. Beispiel:

-----  
ich weis\$, das\$ bei scho\$n Wetter  
-----

#### 18. Klein- und Großschreibung

Außer Substantiven, Eigennamen und der Anrede werden alle Wortformen kleingeschrieben. Dies gilt auch für den Satzanfang.

#### 19. Abkürzungen

In der gesprochenen Sprache gibt es, im Gegensatz zur Schriftsprache, keine Abkürzungen. Entsprechend werden auch solche Wortformen, die üblicherweise Abkürzungen sind, wie gesprochen, aufgeschrieben. Beispiele:

-----  
hier beziehungsweise dort zum Beispiel  
-----

## 20. Stottern

Gestotterte Wörter werden wie nicht voll ausgesprochene Wörter behandelt, d.h. die gestotterten Wortpartikel werden mit einem nachgestellten Bindestrich gekennzeichnet, durch eine Leerstelle voneinander abgesetzt und so oft wiederholt, wie sie hörbar sind. Hier ein Textausschnitt als Beispiel:

-----  
ich ha- ha- hab's doch gleich gesagt  
-----

## 21. Zahlen, Brüche etc.

Zahlen, Brüche etc. werden nach Möglichkeit ausgeschrieben. Lediglich typische Zahlen werden als Ziffernfolge verschriftet. Beispiele:

-----  
elf, zwei Drittel, 1981, James Bond 007, erstens  
-----

## 22. Wortschöpfungen

Werden neue Wörter durch das Zusammenfassen mehrerer einzelner Wörter gebildet, so werden sie mit einem Bindestrich verbunden. Beispiele:

-----  
das In-der-Welt-Sein, schwarz-weiß  
-----

## 23. Fehlleistungen

Versprecher und andere Fehlleistungen werden möglichst wortgetreu in das Transkript übernommen. Beispiele:

-----  
allergierisch, Definieres äh Definiertes  
-----

## 24. Allgemeines

Treffen mehrere Regeln gleichzeitig zu, dann sollte im Interesse der Lesbarkeit der Transkripte versucht werden, mit der Anwendung nur einer Regel auszukommen. Darüber hinaus gelten die Rechtschreibregeln nach DUDEN.

### C Transkriptionsstandard

Zur Transkription von Gesprächen sollten grundsätzlich drei Durchgänge vorgesehen werden, nach deren Abschluß alle Regeln aus den Teilen A und B Anwendung gefunden haben.

Im ersten Durchgang, der eigentlichen Transkription, wird das "Rohtranskript" auf maschinenlesbaren Belegen oder on-line an dem Mikrocomputersystem TELECOMP erstellt. Dabei bleiben die Regeln B2, B10 und B12 ohne Beachtung. Diese erste Fassung des Transkripts wird dem Aufnahmemodul des Textbanksystems zugeführt. Dieser erstellt daraus eine erste Arbeitsfassung und druckt sie in aufbereiteter Form aus.

In einem zweiten Durchgang, dem Kontrollhören, werden auf der ausgedruckten ersten Arbeitsfassung alle Änderungen und Korrekturen handschriftlich vorgenommen. Danach wird das gesamte Gespräch noch einmal ganz abgehört und dabei die Regel B10 zur Markierung der Zeit in Fünf-Minuten-Abständen angewendet. Änderungen und Zeitangaben werden schließlich mit Hilfe eines Korrekturprogrammes direkt in der Textbank vorgenommen.

Der dritte Arbeitsschritt betrifft die Anwendung der Regel B12. Zugrunde liegt eine zweite Arbeitsfassung, die bereits vom Textbanksystem automatisch vorgenommene Zuordnungen gemäß der Mehrdeutigkeitsregel enthält. Über ein interaktives Programm können diese Kodierungen am Bildschirm überprüft und gegebenenfalls geändert werden.

Ein Transkript durchläuft also entsprechend dieser Dreiteilung mehrere Stufen die im folgenden unterschiedlichen Transkriptionsniveaus zugeordnet werden.

Niveau	Beschreibung
0	Es wurden andere oder nur ein Teil der Regeln beachtet.
1	Außer den Regeln B10 für Zeitangaben, B2 für Namen und B12 zur Auflösung von Mehrdeutigkeiten wurden alle Regeln beachtet.
2	Es wurden alle Regeln außer B12 zur Auflösung von Mehrdeutigkeiten streng beachtet.
3	Alle Regeln wurden beachtet.
V	Bei der Transkription wurden nicht alle Regeln beachtet. Es wird jedoch auf dem Begleitzettel festgehalten, welche Regeln nicht verwendet wurden.

Um den Transkriptionsstandard festhalten zu können aber auch für Rückfragen o. ä., wird für jedes zu transkribierende Gespräch ein Begleitzettel geführt. Er enthält unter anderem Angaben zu fol-

genden Punkten:

Name des Transkribenten  
Dauer der Transkription  
Datum der Transkription  
Niveau des Transkripts

Name des Korrekteurs  
Dauer der Korrektur  
Datum der Korrektur  
Niveau des Transkripts

Dauer des Gesprächs

**D Druckaufbereitung der Transkripte**

Transkripte, die auf elektronischen Datenträgern vorliegen und nach den in Teil A und B festgelegten Transkriptionsregeln erstellt sind, können mit dem Programm LAYOUT (siehe Mergenthaler und Stöcklein 1981) aufbereitet und ausgedruckt werden. Das Druckbild eines Textes kann vielfältig variiert und durch eine Seiten-, Zeilen- und Äußerungsnumerierung ergänzt werden. Das Druckprogramm nimmt im Hinblick auf eine bessere Lesbarkeit eine Reihe von Änderungen bei der Druckaufbereitung am Transkript vor. Im einzelnen betrifft dies folgende Regeln:

B1: Die bei der Transkription nachgestellten Zeichen zur Hervorhebung und Dehnung werden in der Druckform den betreffenden Wortformen vorangestellt.

```
-----
:also wenn Sie !das tun
-----
```

Alternativ ist es möglich, die hervorgehobenen Wortformen im Fettdruck und die gedehnten Wortformen gesperrt wiederzugeben.

B2: Der in ein doppeltes Paar runder Klammern eingeschlossene vermutete Wortlaut wird in einfache eckige Klammern eingeschlossen.

```
-----
gestern <vormittag> habe ich
-----
```

B12: Die über einen Schrägstrich an eine Wortform angehängte Auflösung einer Mehrdeutigkeit kann wahlweise beim Ausdruck unterdrückt werden.

```
-----
Bank, wir, dort
-----
```

B18: Nach dem Satzzeichen Punkt wird der erste Buchstabe der nachfolgenden Wortform groß geschrieben.

```
-----
nein= nein. Oder doch? aber
-----
```

**E Hinweise zum Schreiben von Texten****1. Das Erstellen von Belegen**

Belegqualität	holzfreies dichtes Schreibmaschinenpapier, es darf nicht geknickt oder beschmutzt werden.
Belegformat	DIN A4 hoch, allseitiger Rand 2 cm.
Zeilenzahl	28, zweizeilig; entspricht 3 Zeilen/inch.
Spaltenzahl	65, entspricht 10 Zeichen/inch
Farbband	Plastic-Kontrastband wird empfohlen.
Schrifttyp	OCR-A Groß/Kleinschreibung.
Zeichenvorrat	Ziffern: 0123456789 Buchstaben: abcdefghijklmnopqrstuvwxyz ABCDEFGHIJKLMNPOQRSTUVWXYZ Sonderzeichen: /; \$ % ! & * ( ) ? : = ' + " , . -

**Anmerkung:**

Das Ausrufezeichen ist auf dem OCR-A Kugelkopf nicht vorhanden. Als Ersatzdarstellung wird deshalb das Zeichen S (wird auch Fleischerhaken genannt und sieht wie ein eckiges großes S aus) verwendet.

Schriftqualität	Die einzelnen Schriftzeichen müssen kräftig und gleichmäßig gefärbt sein. Ist dies über Anschlags- und Aufschlagstärkeregler nicht erreichbar, muß die Schreibmaschine justiert werden.
Leerzeilen	Leerzeilen dürfen im Text nicht enthalten sein, da sie als Blattende interpretiert werden.
Korrekturen	Einzelne Zeichen können mit dem Delete-Zeichen (schwarz gefärbtes Rechteck) über-tippt werden. Ganze Zeilen können mit einem geraden Bleistiftstrich als ungültig erklärt werden. Korrekturen mit dünn aufgetragenem flüssigem Tipp-Ex sind möglich.
Kennung der Belege	Jeder Beleg und sein Inhalt wird in der ersten Zeile durch eine 14-stellige Nummer gekennzeichnet. Sie muß in Spalte 1 beginnen und sollte keine Korrekturen enthalten.

Die Kennnummer setzt sich aus fünf einzelnen Angaben zusammen, deren Bedeutung nachfolgend erläutert wird.

-----  
 ! ANR ! PNR ! ART ! SNR ! BNR !  
 -----

ANR: dreistellig  
 Nummer für Therapeut, Interviewer, Versuchsleiter o. ä.; wird zentral von der ULMER TEXTBANK vergeben.

PNR: dreistellig  
 Nummer für Patienten, Klienten, Versuchspersonen, Gruppen o. ä.: wird zentral von der ULMER TEXTBANK vergeben.

ART: zweistellig  
 Textart gemäß nachfolgender Tabelle:

- 01 Beratung (D)
- 02 Kurztherapie (D)
- 03 analytische Psychotherapie (D)
- 04 Psychoanalyse (D)
- 05 Paartherapie (G)
- 06 Familientherapie (G)
- 07 Gruppentherapie (G)
- 08 Supportive Psychotherapie (D)
- 09 Gruppenarbeit (G)
- 10 Gesprächstherapie (D)
- 11 Verhaltenstherapie (D)
- 12 Erstinterview - Diagnostik (D)
- 13 Erstinterviewbericht (M)
- 14 Bericht über Psychotherapiestunde (M)
- 15 Bericht über Psychoanalysestunde (M)
- 16 Vorträge allgemein (M)
- 17 Berichte allgemein (M)
- 18 Balintgruppe (G)
- 19 Selbsterfahrungsgruppe (G)
- 20 Träume (D)
- 21 Unterricht (G)
- 22 Psychodiagnostik (D)
- 23 Nachbefragung (D)
- 24 Tat (D)
- 25 Sprachprobe (D)
- 26 Genetische Beratung (G)
- 27 WAT (Wort-Assoziations-Test) (D)
- 28 HIT (HOLZMANN-INKELOT-TEST) (D)
- 29 Erfahrungsbericht (M)
- 30 Wissenschaftliche Abhandlung (M)
- 31 Visitengespräch (G)
- 32 Kognitive Verhaltenstherapie (D)
- 33 Supervision (D)
- 34 Psychiatrische Behandlung (D)
- 99 Sonstiges

Die Buchstaben in Klammern bedeuten:

- (M) = Monolog - ohne Sprecherkennung
- (D) = Dialog - Sprechererkennung \$A und \$P bzw. A/ und P/
- (G) = Gruppengespräch - Sprechererkennung \$A1, \$A2, ... \$P1, \$P2, ... bzw. A1/, A2/, ... bzw. P1/, P2/, ...

SNR: vierstellig  
Nummer für Stunden oder sonstige Texteinheiten.

BNR: zweistellig  
Nummer des Beleges; dient der Kennzeichnung einer Textabfolge Textabfolge.

#### Sprecherkennung

Bei Dialogen werden die einzelnen Sprecher vor Beginn ihres Redebeitrages gekennzeichnet. Mit der Kennung A wird auf die über ANR näher bezeichnete Person, also beispielsweise den Therapeuten und über P auf die über PNR näher bezeichnete Person, also beispielsweise den Patienten hingewiesen. Es sind zwei mögliche Schreibweisen erlaubt: A/ oder \$A bzw. P/ oder \$P. Bei Verwendung des Schrägstrichs muß für die Sprecherkennung mit einer neuen Zeile begonnen werden. Bei Verwendung des Dollarzeichens kann der Sprecherwechsel an jeder beliebigen Stelle einer Zeile festgehalten werden. Allerdings muß in diesem Fall dem Kennbuchstaben A bzw. P eine Leerstelle folgen. Eine gemischte Verwendung ist möglich. Beispiel:

P/letzte Woche war ich aber darüber nicht wütend, hab ich das nicht so gemerkt. \$A  
hmm \$P da ist wahrscheinlich alles...

Bei Gruppengesprächen müssen zur weiteren Unterscheidung einzelner Sprecher bis zu zwei Ziffern unmittelbar an die Kennbuchstaben angefügt werden (z.B. A1/ oder P16/ bzw. \$A1 oder \$P16). Der Sprecherkennung muß eine Leerstelle folgen. Es ist zu beachten, daß bei einem Gruppengespräch immer eine Nummer angegeben werden muß, auch wenn mit der Kennung A oder P nur ein Sprecher beteiligt ist.

Fortsetzungsseite: Die einzelnen Belege eines zusammenhängenden Textes werden über die Blattnummer BNR fortlaufend numeriert. Bei Dialogen oder Gruppengesprächen ist zu beachten, daß die erste Textzeile (d.h. die zweite Zeile des Beleges) mit einer Sprecherkennung beginnen muß. Ist die Äußerung eines Sprechers am

Ende eines Beleges nicht beendet, dann kann auf dem nächsten Beleg fortgesetzt werden, wobei die Sprecherkennung wiederholt wird.

**Belegverarbeitung** Da jeder Beleg durch seine Kennnummer eindeutig identifiziert werden kann, ist für die richtige weitere Verarbeitung keine Sortierung nötig.

**Fehlerhafte Belege** Belege mit Fehlern in der Kennziffer oder einer Sprecherkennung werden mit einem entsprechenden Fehlerhinweis angesteuert und können nach entsprechenden Korrekturen nochmals eingelesen werden. Folgende Meldungen sind möglich:

\*1 14 BN Die Blattnummer in der 1. Belegzeile ist nicht 14-stellig.

\*2 NN BN Die Blattnummer enthält ein nichtnumerisches Zeichen.

\*3 RJ FS Zxx In Zeile xx ist mindestens ein unkenntliches Zeichen.

\*4 UBRGR Die letzte Belegzeile ist zu dicht am unteren Blattrand.

1 F89 PKIL In der ersten Textzeile eines Beleges fehlt am Zeilenanfang eine Sprecherkennung.

2 FA PK Z.xx Die Personenkennung in Zeile xx ist falsch.

3 PKZL Z.xx Die Personenkennung in Zeile xx ist falsch.

UNDE FEME Nicht genauer beschreibbarer Fehler.

Die mit dem Zeichen \* gekennzeichneten Meldungen können auch durch eine Verschmutzung der Belege verursacht sein.

## 2. On-Line-Transkription am TELECOMP

Die Transkription am TELECOMP wird durch ein spezielles Programm unterstützt. Nachdem der Benutzer sich unter dem Namen TRANS angemeldet hat, werden alle weiteren benötigten Angaben von ihm im Bildschirmdialog erfragt. Anschließend kann der Text eingegeben werden.

## F Übersicht zur Verwendung der Zeichen

Zeichen	Verwendung	Beschreibung	Seite
?	Satzzeichen	B 5	4
	Interjektion	B 11	7
=	Satzzeichen	B 5	4
	Interjektion	B 11	7
,	Satzzeichen	B 5	4
	Interjektion	B 11	7
.	Satzzeichen	B 5	4
	Interjektion	B 11	7
;	Satzzeichen	B 5	4
/	Unverständliche Wortform	B 8	5
	Mehrdeutigkeit	B 12	7
*	Ersatznamen	B 2	3
-	Wortabbrüche	B 7	5
	Zeitintervalle bei Pausen	B 9	6
	Diskontinuierliche Formen	B 15	8
	Stottern	B 20	9
	Wortschöpfungen	B 22	9
( )	Kommentare	B 13	7
	Nicht verbale Äußerungen	A 3	2
	Situationsgebundene Geräusche	A 4	2
	Zeitangaben	B 10	6
(( ))	Vermuteter Wortlaut	B 8	5
'	Zitate	B 3	3
	Minuten (bei Zeitangaben)	B 10	6
"	Sprachwechsel	B 4	3
	Sekunden (bei Zeitangaben)	B 10	6
+	Gleichzeitigkeit	B 6	4
	Zeitangaben	B 10	6
:	Dehnung	B 1	3
!	Hervorhebung	B 1	3
< >	Vermuteter Wortlaut (nur in der Druckform)	D	12
§	Worttrennung am Zeilenende	B 14	7
&	ohne besondere Bedeutung		

**Literaturhinweise**

Bausch, Karl-Heinz: Zur Umschrift gesprochener Hochsprache. In: Steger, Hugo (Hg.): Texte gesprochener deutscher Standardsprache I (S. 33-54), Institut für Deutsche Sprache, Freiburg 1971.

Drewek, Raimund: Codierungsregeln für Texte. In: Dokumentation der LDVLIB-Programme für computerunterstützte Textanalysen, Stuttgart 1980.

Ehlich, Konrad und Rehbein, Jochen: Halbinterpretative Arbeitstranskriptionen (HIAT). Linguistische Berichte 45 (21-41) 1976

Ehlich, Konrad und Rehbein, Jochen: Erweiterte halbinterpretative Arbeitstranskriptionen (HIAT2). Linguistische Berichte 59 (51-75) 1979

Ehlich, Konrad und Switalla, Bernd: Transkriptionssysteme - Eine exemplarische Übersicht. Studium Linguistik (78-105) 1976

Henne, Helmut und Rehbock, Helmut: Einführung in die Gesprächsanalyse. Sammlung Göschen, Walter de Gruyter Berlin 1979 (S. 39-157).

Klann-Delius, Gisela: Zur Transkription der Daten. Arbeitspapier. Projekt: Erwerb von Diskursfähigkeit, Bielefeld 1981

Mergenthaler, Erhard: Das Textkorpus in der psychoanalytischen Forschung. In: Bergenholtz, Henning und Schaefer, Burkhard (Hg.): Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora, Scriptor, Königstein/Ts. 1979 (S. 131-147)

Mergenthaler, Erhard und Stöcklein, Ursula: Der LAYOUT-Modul. EU 1/1981. SFB 129 Projekt B2. Universität Ulm 1981

Switalla, Bernd: Die Identifikation kommunikativer "Daten" als sprachtheoretisches Problem. Semiotik 1 (161-175) 1979

Wenzel, A.: Zum Verhältnis von verbaler und nonverbaler Kommunikation mit einer exemplarischen Analyse eines Gesprächsausschnitts. deutsche sprache (1-20) 1980

Winkler, Peter: Notationen des Sprechausdrucks. Semiotik 1 (211-224) 1979



TEXTAUFBEREITUNG  
FÜR DIE  
EDV-UNTERSTÜTZTE TEXTANALYSE  
IN DEN  
SCZIALWISSENSCHAFTEN

Peter Ph. Mohler

ZUMA Arbeitspapier  
Nr. 82/05  
Mai 1982

ZUMA  
B 2, 1  
68 Mannheim

Inhalt

	Seite
1. Textverarbeitung vs Textanalyse	1
2. Textaufbereitung für die Textanalyse	1
3. Probleme der Textaufbereitung bei direkter Eingabe in den Computer	4
4. Beispiel einer Transkription mittels direkter Dateneingabe	7
5. Zusammenfassung	15
6. Anhang Verschriftungsbeispiel	17

## 1. Textverarbeitung vs. Textanalyse

"Textverarbeitung" ist gegenwärtig ein wichtiger Zweig kommerzieller Datenverarbeitung. Dieser Begriff ist ein Euphemismus für den älteren Begriff der alpha-numerischen Datenverarbeitung und zugleich dessen Einschränkung. Er meint die Automatisierung von Büroschreibaarbeiten aller Art. Also vom Geschäftsbrief über das Formular, Adressenaufklebern zu Druckvorlagen für Offsetdruck. Andere Teile der alpha-numerischen Datenverarbeitung wie z.B. die Automatisierung des Rotationsdruckes oder die Dokumentation von Bibliotheksbeständen gehören nicht zu diesem Bereich, auch wenn es offensichtliche Überschneidungen in der Programmlogik gibt. Man kann sagen, daß Textverarbeitungsverfahren im wesentlichen die Textproduktion (in Büros) unterstützen.

Textanalyseverfahren hingegen unterstützen die Auswertung von Texten. Dafür müssen solche Texte EDV-gemäß aufbereitet sein. Dies kann schon bei der Produktion des Textes wie z.B. bei Zeitungen mit automatisiertem Rotationsdruck geschehen (De Weese, 1974). Für die Textanalyse müssen solche Vorlagen dann nur noch umformatiert werden; Druckwerke können mit Omnifont-Lesern in EDV-Dateien umgesetzt werden (Kurzweil, 1981). Zumeist müssen aber die zur Analyse vorgesehenen Dokumente neu in einem EDV-gerechten Format und Medium abgeschrieben werden.

## 2. Textaufbereitung für Textanalyse

Bei ZUMA werden zur Zeit Texte mit dem OCR-Verfahren edv-gemäß aufbereitet. Dabei wird ein Text in einer Spezialschrift auf Papier geschrieben und dann mit einem optischen Leser auf Magnetband übertragen.

Das Verfahren hat sich bewährt, leidet aber unter folgenden Nachteilen:

- a ) zur Zeit ist Groß-Klein-Schreibung nicht möglich. Da die neuen Textanalyseprogramme (EVA und TEXTPACK V) diese berücksichtigen und damit wesentliche Erleichterungen für die Analyse gegeben sind, erweist sich dieser Nachteil als erheblicher methodischer Mangel.
  
- b ) Es lohnt sich aus Kostengründen nur eine größere Menge gleichartigen Materials an die auswärtige OCR-Lesemaschine weiterzugeben. Dies hat folgende Nachteile:
  1. in der Erprobungsphase von Transkriptionen aller Art, sollten unterschiedliche Transkriptionsregeln auf ihre Nützlichkeit für den Analysezweck hin erprobt werden. Dies ist zur Zeit nur unter Inkaufnahme erheblicher Kosten und Zeitverzögerungen (Zeit: ca. 3 - 4 Wochen nach der Transkription!) möglich.
  2. Transkriptionen großer Mengen Text sind in der Regel sehr zeitaufwendig. Die Transkription großer Texte kann sich über mehrere Monate hinziehen. Beim OCR-Verfahren geht in der Regel diese Zeit für die Analysevorbereitungen (z.B. Entwicklung eines Klassifikationsschemas) verloren.

Eine weitere Möglichkeit ist die direkte Eingabe von Texten in einen Computer.

Diese direkte Dateneingabe vermeidet die o.a. Nachteile des OCR-Verfahrens. D.h. Groß-Kleinschreibung ist möglich, und jeder beliebig kleine Textabschnitt steht nach der Eingabe und Korrektur sofort für die Analyse zur Verfügung. Dadurch wäre die Möglichkeit sukzessiver Entwicklung von Kategoriensystemen bzw. Klassifikationsschemata wie sie von Namenwirth gefordert wird gegeben (Namenwirth et al. 1978).

### 3. Probleme der Textaufbereitung bei direkter Eingabe in den Computer

Geringe Schreibgeschwindigkeit, Ausfallzeiten der EDV-Anlage, schnellere Ermüdung der Schreibkraft am Bildschirm, psychologische Sperrren gegen EDV-Arbeit und Dauer bei der Trainingsphase für die Schreibkraft können wesentliche Nachteile der direkten Texteingabe sein.

#### a ) geringe Schreibgeschwindigkeit

Bei direkter Eingabe in eine Groß-Rechenanlage ist es möglich, daß, je nach momentaner Auslastung des Computers, die Antwortzeiten für professionelle Schreibkräfte zu lange sind. "Antwortzeiten" bedeutet die Zeit zwischen dem Absenden einer Bildschirmzeile an den Computer bis zur Quittierung dieser Zeile durch den Computer. Erst nach der Quittierung kann weiterer Text eingegeben werden. Praktisch kann man dies mit der Wagenrücklaufzeit einer gewöhnlichen Schreibmaschine vergleichen. Es ist nicht ungewöhnlich, daß die Antwortzeiten bis zu mehreren Sekunden, in extremen Fällen mehrere Minuten dauern können!

Dies Problem erübrigt sich, wenn die Eingabe über eine, von der Zentraleinheit unabhängige Zwischenstation (z.B. intelligenter Bildschirm mit eigenem Speicher) oder auf einen eigenen Rechner, dessen Auslastung vom Forscher kontrolliert werden kann, erfolgt.

#### b ) Ausfallzeiten der EDV-Anlage

Wartungsarbeiten und die Dauer bzw. den Zeitpunkt anderer Ausfallzeiten können bei Hochschulgroßrechnern vom Forscher nicht bestimmt werden. Er ist dabei voll auf die Hilfe des Hochschulrechenzentrums angewiesen.

Diese Ausfallzeiten können folgende Ursachen bzw. Wirkungen haben:

Wenn die Zentraleinheit nur zu den üblichen Bürozeiten gewartet bzw. repariert wird, so ist der Zeitraum für professionelle Dateneingabe, die eben zu Bürozeiten stattfindet, eingeschränkt.

Wartung und Reparatur sind allerdings kleinere und vor allem vorhersehbare (Wartung) oder hinzunehmende (Reparatur) Übel. Erschwerender ist die Angewohnheit einiger Rechenzentren Tests & Umstellungen etc. unangemeldet und zu Bürozeiten durchzuführen.<sup>(1)</sup>

Datentypistinnen, die keine anderen Aufgaben sonst zu erfüllen haben, müssen für diese Ausfallzeiten natürlich bezahlt werden.

Diese organisatorischen Schwierigkeiten, die periodisch und dann zumeist gehäuft auftreten, können eine ernsthafte Behinderung eines wissenschaftlichen Projektes sein.

Wiederum kann dieses Problem durch eigene Datenaufzeichnungsgeräte umgangen werden.

c ) schnellere Ermüdung durch Bildschirmarbeiten

Alle, die mit Bildschirmen und mit konventionellen Schreibmaschinen arbeiten, berichten von stärkerer Ermüdung durch die Bildschirmarbeit als durch konventionelle Schreibarbeiten.

---

(1)

z.B. Kontakt zur Zentraleinheit bricht ab, Anruf beim Operator und dessen Hinweis "Ach ja, heute probieren wir eine neue Schaltung aus".

Erste Ursache sind die, zumindest im universitären Bereich vorhandenen, unzureichenden Arbeitsplätze, d.h. Stuhl, Tisch, Beleuchtung (!) und Bildschirmgerät müssen aufeinander abgestimmt sein und den Erfordernissen professioneller Schreibkräfte entsprechen.

Zweite, nicht behebbare Ursache, ist die größere Konzentration, die die Bildschirmarbeit offensichtlich von den Schreibkräften abverlangt. Dies zu erörtern, würde hier zu weit führen, fest steht, daß es unmöglich scheint, am Bildschirm "unterbewußt" zu arbeiten.<sup>(1)</sup>

Es ist deshalb zu prüfen, ob nicht sogenannte Blattschreiber, also an den Computer angeschlossene Schreibmaschinen für die Texterfassung geeigneter sind als Bildschirme.

d ) psychologische Sperrn gegen EDV-Arbeiten

Für Projekte kann es schwierig werden, Personen für die Transkription zu gewinnen, die schnell schreiben können und bereit sind mit EDV-Anlagen zu arbeiten.

Die psychologischen Barrieren sind nicht unbegründet, da oberflächlich vergleichbare Arbeiten, wie z.B. Transkriptionen von Texten, aus vielfältigen Gründen tiefgreifende Unterschiede zwischen konventionellen und EDV-Arbeiten zeitigen. Die Symbolisierung von Arbeitsschritten in der EDV (die Maschine wird z.B. nicht "eingeschaltet" sondern mit Symbolen "aktiviert", die Zeilenbreite wird nicht am Randsteller "eingestellt" sondern mit Symbolen programmiert) und die geringere, vielleicht nur empfundene, Autonomie des Schreibers sind wohl die wichtigsten Gründe.

---

(1) damit ist das Phänomen gemeint, daß professionelle Schreibkräfte untereinander Gespräche führen können, während sie Vorlagen abschreiben.

#### e ) Längere Trainingsphase

Die psychologischen Barrieren bedingen eine längere Trainingsphase der Schreibkräfte, weil sie nicht nur die Transkriptionsregeln sondern auch den Umgang mit einer, ihnen manchmal unheimlichen, Maschinerie lernen müssen.

#### 4. Beispiel einer Transkription mittels direkter Dateneingabe

Für das Projekt "Wohlfahrtsansprüche" (82/01) wurde bei ZUMA die direkte Transkription von Tonbandmitschnitten (Interviews) erprobt.

Der Test wurde an einem Bildschirm, der an einen Siemens-Rechner unter BS 2000 angeschlossen ist, durchgeführt. Die Eingabe wurde vom Programm EDT unterstützt.

##### 4.1 Durchführung / Programm

Der Transkribent sollte soweit als möglich von Kenntnissen über den Computer in der Anfangsphase entlastet werden, da die Vielzahl kleinster, für den Anfänger nicht zusammenhängender Informationen dem Ziel einer raschen Einarbeitung in die Transkription (und nicht in EDV-Arbeiten!) entgegensteht.

Deshalb wurden für den EDT Hilfsprozeduren geschrieben, die bei Dateiverwaltung, also Laden und Sichern der geschriebenen Texte und Zeit/Datumsangaben steuern. Die Zeit/Datumsangaben erlauben Berechnungen für den Zeitaufwand einzelner Transkriptionen.

Folgende Arbeitsschritte mußte der Transkribent lernen:

- a) Ein/Ausschalten des Bildschirms
- b) Aktivieren des Bildschirms:  
 Funktionstasten zum Anmelden an das Betriebssystem.  
 Datenübertragungstasten  
 LOGON-Kommando  
 auf die Anfrage CONTINUE (NEWS): no.
- c) Editoraufruf  
 §EDT
- d) laden der Datei und der Hilfsprozeduren im Editor  
 @ INPUT 'PM.8201.INPUT' (zugleich Datum und Zeit)
- e) Block-Mode C übertragen ganzer Bildschirmseiten:  
 @ BL ON
- f) Tabulatorsymbol (ein Symbol, das, wenn es übertragen wird, wie die Tabulatortaste einer Schreibmaschine funktioniert) hier wurde ] (eckige Klammer) gewählt.
- g) Schreibregeln  
 nach Satzzeichen keine Großschreibung.  
 Umlaute, ß etc.: ae, ue, ss etc.  
 Identifikationsnummern vergeben; hier Interviewnummer, Interviewer/Befragter/Transkribent.
- h) Zwischensichern  
 Abspeichern des Geschriebenen nach unregelmäßigen Zeitabständen zur Zwischensicherung der Datei (zugleich festhalten der Zeit).  
 @ DO 1
- i) zusätzliche Kommandos im EDT:  
 @ P a-b Ausgabe von Zeilen auf den Bildschirm.

Erläuterungen zu den EDT-Prozeduren:

- ad d): @ INPUT 'PM.8201.INPUT'  
 dies File 'PM.8201.INPUT' hat folgenden Inhalt:

BEISPIEL EINER EDT PROZEDUR FÜR DIREKTE DATENEINGABE

Diese Prozedur wird mit dem Kommando 'input' in den Editor geladen und erzeugt dort zwei Unterprozeduren (Proc 1 und Proc 2)

```

*DJWER ON          - Groß- Kleinschreibung
*CM ON            - check on: Änderungen werden quittiert
*SET 'TEXTDATEI' - lade Textdatei in den Editor
*TAGS : :I:8      - Tabulatorzeichen = Sprung in Spalte 8
*SET #L1=*
*SET #S1=DATE          DATUM: ',#S1
*CR #L1:'***
*SET #L1=**+1
*SET #S2=TIME          UHRZEIT: ',#S2
*CR #L1:'***
*PRC 1
**SA 'TEXTDATEI'UPDATE
**SET #L1=**+1
**SET #S2=TIME          UHRZEIT: ',#S2
**CR #L1:'***
**R
**RETURN
**END
*PRC 2
**SET #L1=**+1
**SET #S2=TIME          UHRZEIT ENDE: ',#S2,' DATUM: ',#S1
**CR #L1:'***
**SET #L1=#L1**+1
**CR #L1:'***
**SA 'TEXTDATEI'UPDATE
**SY*PR TEXTDATEI,T=STATIONZ,LINES=60
**SY'LOGOFF
**END

```

Datum und Uhrzeit bei Anfang der Dateneingabe werden in die Textdatei geschrieben und auf dem Bildschirm ausgegeben

Proc 1 : Prozedur zur Zwischensicherung der Eingabe

Proc 2 : Ende der Eingabe

ENDE DER EINGABE \*\*\*\*\*

## 4.2 Ergebnisse

Für die o.a. Probleme ergab sich folgendes:

### a) Schreibgeschwindigkeit

für die Transkription von Tonbandmitschnitten, bei denen die mögliche Schreibgeschwindigkeit an sich relativ langsam ist, war die Antwortzeit dann kurz genug, wenn nicht jede Bildschirmzeile sondern ganze Bildschirmseiten übertragen werden (Block-Mode).

### b) Ausfallzeiten der EDV-Anlage

Die Transkription wurde von einer Person durchgeführt, die auch andere Aufgaben zu erledigen hat. Insofern ergaben sich keine Probleme mit Ausfallzeiten.

### c) Ermüdung

Zu Beginn der Transkription, in der Trainingsphase, betragen die Verweilzeiten des Transkribenten am Bildschirm ca. 20 Minuten bis 1 Stunde. Später bis zu zwei Stunden.

Die Eingabe wurde durch die ungenügende Arbeitsplatzgestaltung, insbesondere durch die miserable Tastatur des Bildschirms erheblich behindert.

### d) psychologische Sperren

Obwohl der Transkribent schon an einem Mikrocomputer (Betatext) Texteingaben und - Korrekturen in größerem Umfang machte, waren seine Abneigungen gegen die neue Arbeit erheblich. Dies beruhte in großem Maße auf Ängsten vor dem neuen Gerät (an dem sonst nur "Wissenschaftler" arbeiten) und wahrscheinlich auf der Furcht, zu versagen.

e) Trainingsphase

Das Training verkürzte sich erheblich durch die festen Prozeduren, die dem Transkribenten von der Dateiverwaltung entlasteten. Es dauerte ca. 2 Wochen, bis der Transkribent sich "heimisch" fühlte, soweit das mit EDV überhaupt möglich ist.

4.3 Bewertung

a) Zeitaufwand gegenüber konventioneller Verschriftung

Die Dauer der Transkription an sich verkürzt sich nicht durch die direkte Dateneingabe. Die Erfahrungswerte der ZUMA-Feldabteilung liegen beim 6 - 10 fachen der gesprochenen Zeit, diese wurden beim Test eingehalten. Nachteilig für die direkte Dateneingabe ist die längere Einarbeitungsphase, die aber auch bei der OCR-Verschriftung zu beobachten ist!

b) Genauigkeit

Ein spezieller Genauigkeitstest wurde nicht durchgeführt. Allerdings zeigte sich beim Vergleich einer EDV - zu einer Normal-Transkription, daß die Möglichkeit des Transkribenten, unter einer eigenen Identifikationsnummer beliebige Kommentare einzufügen, als sehr hilfreich für das Verständnis des Lesers. Gleichzeitig können so vom Transkribenten Markierungen für die Stellen gesetzt werden, an denen er seiner Transkription nicht sicher war oder nicht transkribieren konnte. Diese Stellen können dann leicht vom Korrektor herausgesucht werden und gegebenenfalls verbessert oder ergänzt werden.

Da auch die Korrekturschritte mit Kommentaren, die in der Analysephase ignoriert werden, versehen werden können, wird die Genauigkeit von direkt eingegebenen Texten insgesamt größer sein, als die konventioneller Verschriftung.

Man muß dabei allerdings berücksichtigen, daß letzteres, die Dokumentation von Korrekturen, auch bei OCR und anderen Arten der EDV-Aufbereitung von Texten möglich ist.

### c) Analysephase versus Aufbereitungsphase

Im Gegensatz zu allen anderen EDV-Aufbereitungsarten für Texte (wie auch für numerische Informationen) erlaubt nur die direkte Dateneingabe eine wirtschaftlich vertretbare zeitliche Überlappung von Datenerfassung, Datenkorrektur und Datenanalyse. Diese drei Phasen werden in der Regel zeitlich voneinander getrennt gehalten. Bei der langen Zeit, die bei den Textdaten für die ersten beiden Phasen benötigt wird, ist dies nicht sinnvoll.

Der Ablauf eines Projektes, das sich direkter Dateneingabe bedient wird nun skizziert:

Noch während im Feld Daten erhoben werden (Interviews) werden die ersten eingegangenen Daten erfaßt. Für eine Stunde Interview sind, abhängig von der Kommunikationsstruktur des Interviews und der Aufnahmequalität, sechs bis zehn Stunden Transkription anzusetzen, d.h. unter optimalen Verhältnissen (Training und Arbeitsplatz) kann ein Interview von einer Stunde Dauer in ein- bis zwei Tagen transkribiert werden, wenn man dafür eine Volltagsschreibkraft einsetzt.

Danach beginnt sofort die erste Datenkorrektur, d.h. ein anderer Mitarbeiter hört das Interview ab, vergleicht den Text und verbessert ihn gegebenenfalls. Diese zweite Korrektur dauert pro Interviewstunde ca. 2 - 4 Stunden. Nach den bisherigen Erfahrungen ist solch eine Korrektur dringend notwendig, da sich beim Transkribieren systematische Fehler, wie z.B. Wortersetzungen einschleichen. In den folgenden Korrekturschritten wird dann je nach Zielrichtung des Projektes, eine Feinkorrektur und Feinvercodung durchgeführt (z.B. die Transkription / Vercodung von Pausen, Hintergrundgeräuschen, Intonation etc.) (Tillmann, 1981)

Damit ist ein gleitender Übergang zur Analysephase gegeben, in der jetzt Klassifikationsschemata oder andere Ordnungskriterien entwickelt und auf Teile bzw. schließlich den gesamten Text angewendet werden.

Nimmt man als Beispiel nun ein Projekt mit hundert einstündigen Interviews so ergeben sich überschlägig folgende Zahlen:

Feldzeit: im günstigsten Fall	1 Monat
Transkription: (100 x 6 Std. wenn 1 Person transkribiert)	4 Monate
1. Korrektur	1 Monat

D.h. bei völliger Trennung der drei Phasen verginge ein halbes Jahr, bevor mit der Analysephase begonnen werden könnte.

Unter besten Bedingungen, d.h. 1 Person, die vollzeitleich transkribiert, 1 Person die halbtags korrigiert und 1 - 2 Personen, die das Klassifikationsschema entwickeln könnten, bei direkter Dateneingabe und Ausnutzung aller Rationalisierungsmöglichkeiten der EDV, bei



dem von uns vorgeschlagenen Verfahren nach sechs Monaten erste Ergebnisse vorliegen. Bei der üblichen Projektplanung (1 Jahr Vorbereitung, 1 Jahr Datenerhebung und Analyse), wäre dann noch genügend Zeit für tieferegehende Analysen.

d) Möglichkeiten nicht-quantitativer Analysen

Die Möglichkeiten quantitativer Textanalysen mittels EDV (computerunterstützte Inhaltsanalyse, cui) sind an anderer Stelle beschrieben (Mochmann 1978)

Auf die Möglichkeit cui auch für nicht-quantitative Analysen einzusetzen soll hier kurz hingewiesen werden. Eine längere Erörterung dieses Gegenstandes ist in Vorbereitung.

Ein Problem aller Textanalysen, ist die Organisation der Texte, insbesondere die Permutation der Zitate zu gewissen Kategorien, Eigenschaften (Zettelkasten).

Der Aufbau dieses Zettelkastens, dessen Permutation enge arbeitsökonomische Grenzen gesetzt sind, kann mit cui-Verfahren erleichtert werden.

Hier nur eine Liste möglicher Vorgehensweisen:

1. Auffinden bestimmter Worte, Wortketten  
im Text: Key Word in Context, Search im EDT.
2. Markieren / Kodieren von Textstellen, anstelle von  
abschreiben und wiederfinden der Markierung  
wie unter 1.
3. Kodierung äußerer (z.B. Verlag, Autor etc.)  
oder innerer (z.B. Markierungen wie unter 2)  
Merkmale und Selektion von Zitaten nach bestimmten  
dieser Merkmalskombinationen.

4. Den Text nach bestimmten Merkmalen sortieren  
(z.B. nach Endungen).

5. Zusammenfassung

Die direkte Eingabe von Texten in Computer kann Forschungsprojekte, die größere Mengen solcher Daten bearbeiten wollen, wesentlich entlasten.

Der Vorteil liegt vor allem im Zeitgewinn gegenüber anderen Verfahren der Transkription. Um diese Vorteile ausnutzen zu können müssen die in Teil 3. angeführten Probleme, die allerdings vorwiegend organisatorische sind, gelöst werden.

Bibliographie

- De Weese, L.G., 1977, Computer Content Analysis of 'day-old' newspapers: a feasibility study, Public Opinion Quarterly 41:91-94.
- Kurzweil/Omnifont, 1981, Prospekt, Kurzweil Computer Products, 33 Cambridge Parkway, Mass. 02142.
- Namenwirth, Zvi, et.al., In search of semantic characteristics for machine coding, in: Methoden en Data Nieuwsbrief, 1, Febr. 1978, 3. Jahrgang, S.75-88, Amsterdam.
- Tillmann, H.G., Über den Gegenstand der phonetischen Transkription, in: P. Winkler (Hrsg.), Methoden der Analyse von Face-to-Face Situationen, Stuttgart, 1981, S. 56-62
- Mochmann, E. (Hrsg.), Computerstrategien in der Kommunikationsanalyse, Frankfurt, 1978.

Anleitung zum Direct  
Data Entry

Blanca Degenhardt

Tel.: 089/2180-2250

Blanca Degenhardt

ANLEITUNG ZUM 'DIRECT TEXT DATA ENTRY'

Liebe Schreibkraft,

Du nimmst an einem Methodenexperiment teil, das den Einsatz des Computers bei der Inhaltsanalyse von Leitfadeninterviews zum Inhalt hat.

Um Dir die Texteingabe so leicht wie moeglich zu machen, haben wir eine Eingabeprozedur geschrieben, die es Dir erspart, sich mit dem Betriebssystem des Rechners naeher befassen zu muessen. Trotzdem gibt es noch verschiedene Dinge zu beachten, die wir im folgenden zusammengestellt haben:

1. Allgemeines zur Texteingabe:

Der Text wird fortlaufend von Band geschrieben. Du brauchst Dich nicht mehr um die Seitenlaenge zu kuermern. Diese rechnet die Eingabeprozedur automatisch nach jeder Sitzung aus. Du muessst auch nicht mehr auf jede Seite die Fallnummer und den Zeitpunkt schreiben. Diese Arbeit nimmt Dir der Computer ebenfalls ab.

Nach jeder Zeile muessst Du auf die 'break/send'-Taste (im folgenden mit b/s abgekuerzt) druecken, um zur naechsten Zeile weiterspringen zu koennen. Nie ueber das Zeilenende hinauspringen!

Die Einteilung des Interviews folgt den ueblichen Konventionen:

F: fuer den Interviewer (Frage)  
L: fuer den Interviewten (Lehrer)

2 Leerzeichen nach dem ':', die Folgezeile beginnt am

Anfang der Zeile.

Bei P a u s e n im Interview gib bitte in Klammern an "Pause" und verbalisiere die Laenge der Pause, z.B.:

(kurze Pause, stockt, o.ae.)

Sonstige Aeusserungen, wie Lachen, Raeusperrn usw. gib bitte wie ueblich in Klammern an.

Bei unvertaendlichen Suetzen bzw. Satzteilen gib bitte in Klammern an "unvertaendlich" und dahinter die ungefaehre Laenge der Suetze mit .... an, z.B.:

(unvertaendlich, .....)

U m l a u t e werden wie folgt geschrieben:

ae : \$a

ue : \$u

oe : \$o

Dies gilt fuer Gross- und Kleinbuchstaben.

'Scharfes ss' : \$s

83/06/26. 21.55.18. AISZ08S  
LRZ CYBER175-AA00 MULTI MF.  
FAMILY: ,ub221aa,petra  
TERMINAL: 6, NAM:IAF  
RECOVER/ CHARGE: charge,\*

NOS531AA00/830317.

	CHARGE	PROJEKT
GUTHABEN (SRUS)	2327	
ZUWACHS (SRUS/TAG)	50	

\*\*\*\*\* L R Z - K U R Z M I T T E I L U N G \*\*\*\*\*

21.06.83 12 UHR 00 WALTER  
NEUES SATTELRUNDSCHREIBEN:  
ZUGRIFF UEBER  
INFO,RUNDSCH,83010B

ENTHAELT U.A. WICHTIGE INFORMATIONEN UEBER  
BEUTELFURSE  
PSI-80 MIT DOPPELTER SCHREIBDICHT  
TESTCOMPILER WERDEN ZURUECKGEZOGEN  
NEUE VERSIONEN VON PROSEM UND LISP  
ABLOESUNG VON SYMBOL 4.

SETTL , 520 .THE LIMIT  
/-intext,,iris

— Beginn der Texteingabe  
—

? Jetzt kann mit der Texteingabe begonnen werden.

?

? Wenn Du aufhören willst, mußt Du zuerst eine leere Eingabe machen:

?

?

?? Jetzt erscheinen 2 Fragezeichen !!!

?? Dann einfach "c" eingeben:

INPUT

?

EDIT

?? q

— Ende der Texteingabe

REVERT. \*\* ENDE VON IRIS

/bye

UB221AA LOG OFF 22.00.50.

UB221AA SRU 1.140 UMIS.

IAF CONNECT TIME 00.04.39.  
LOGGED OUT.

### 3. Korrekturen

Wenn Du Dich in einer Zeile vertippt hast, kannst Du den Fehler nur dann verbessern, wenn Du noch in der Zeile bist, d.h. wenn Du noch nicht auf die b/s-Taste gedrueckt hast. Welche Korrekturmoeglichkeiten Du in der Zeile hast, zeigen wir Dir bei der praktischen Einweisung.

Entdeckst Du einen Tippfehler, nachdem Du bereits weitergesprungen bist, kannst Du keine Korrekturen mehr einfuegen.

Faellt Dir auf, dass Du in den Zeilen vorher falsch geschrieben hast (Worte vergessen oder veraendert), trage bitte die korrekten Zeilen noch einmal ein und mach eine Anmerkung, z.B.:

<falsch>: L: Ich hab da so nen Kadett .....  
<richtig>: (KORREKTUR)  
L: Ich hab da so nen alten Kadett ...

#### 4. Fehlermöglichkeiten

Ausser Schreib- und Hoerfehler gibt es leider beim 'direct text data entry' noch andere Fehlermoeglichkeiten:

- FEHLER:** Druecken der Taste b/s direkt nach '?', ohne Leerzeichen dazwischen; bewirkt, dass Du aus dem Eingabemodus herausgehst. Es erscheint: '??'
- ABHILFE:** Du drueckst die b/s-Taste noch einmal und bist wieder im Eingabemodus, es erscheint '?' und Du kannst direkt weiterschreiben
- FEHLER:** Druecken der Taste b/s aus Versehen zweimal. Es erscheint: 'TAKE 80 CHARACTERS', manchmal piept's auch dazu.
- ABHILFE:** Loeschen der gesamten Bildschirmseite, dann weiterschreiben, aber die zuletzt eingegebene Zeile noch einmal schreiben.
- FEHLER:** Druecken der grauen 'LESS'-Taste
- ABHILFE:** Diese Taste darf unter keinen Umstaenden gedruickt werden, weil dies zum Verlust des gerade geschriebenen Textes fuehrt. Du musst deshalb die Taste mit einem Stueck Papier abdecken.
- FEHLER:** Knoten- bzw. Rechnerabsturz, am Bildschirm erscheint: 'HOST UNAVAILABLE'
- ABHILFE:** Je nach Schwere des Absturzes kann es bis zu einer halben Stunde dauern, bis der Rechner wieder arbeitet. Falls es Dir zu lange dauert, geh ruhig nach Hause, ruf mich aber bitte gleich an, damit ich die Texte sichern kann. Der gerade geschriebene Text geht auf keinen Fall verloren.
- Falls Du jedoch warten willst, erscheint nach einiger Zeit:

'BEST AVAILABLE'  
und dann die uebliche Abfrage nach den  
Benutzerkennzeichen (FAMILY usw.)  
Achtung: Es dauert zwischen den einzelnen  
Anfragen meist einige Minuten,  
nicht ungeduldig werden!

Wird die Anfrage 'RECOVER/CHARGE' aus-  
geschrieben, antworte bitte jetzt mit:

RECOVER

und druecke die b/s-Taste.

Darauf erscheint:

'TO CONTINUE ENTER CR'

und Du drueckst wieder die b/s-Taste  
(CR bedeutet 'Carriage Return' und be-  
zeichnet die b/s-Taste).

Wenn alles gut geht, bist Du jetzt wieder  
in alten Eingabemodus und Du kannst weiter-  
schreiben, wo Du aufgehoeert hast.

Es empfiehlt sich jedoch,, die zuletzt ein-  
getragene Zeile noch einmal zu wiederholen.

Manchmal ist ein 'RECOVER' jedoch nicht mehr  
moeglich und der Rechner gibt die entsprechen-  
de Meldung:

'RECOVERY IMPOSSIBLE'

Du kannst jetzt nicht mehr weiterschreiben,  
schalte bitte ab und ruf mich sofort an.

Der gerade eingetragene Text ist nicht ver-  
lorengegangen.

## 2. Textmanipulationen und Textanalysen

Die drei folgenden Beiträge befassen sich mit einigen grundlegenden Manipulations- und Analysemöglichkeiten von Texten, die in Computern gespeichert sind. In SAR (Simple And Retrieval) beschreibt P. Mohler, wie mit Hilfe des Siemens EDT-Programmes Textstellen (Zeilen) gefunden werden können, in denen zwei bestimmte Wörter zusammen auftreten.

Im folgenden kurzen Ausschnitt aus dem TEXTPACK V-Handbuch, werden einige Möglichkeiten dieses Programms und ein für die Textanalyse notwendiger Teil des Computerrotwelsch beschrieben.

J. Schupps Beitrag schließlich zeigt, wie mit verhältnismäßig geringem EDV-Aufwand, sich erhebliche Erleichterungen der qualitativen Textanalyse erreichen lassen.

S A R

Simple And Retrieval

mit dem Siemens Texteditor EDT

Peter Ph. Mohler

ZUMA Arbeitspapier

Nr. 13/82

Juli 1982

ZUMA

B 2,1

68 Mannheim

## 1. Einleitung

Textverarbeitung mit Computern, gleich ob es sich um computerunterstützte Inhaltsanalyse oder eher qualitativ ausgerichtete Forschungsansätze handelt, ist oftmals mit "Information Retrieval" (Informations-Rückgewinnung) verbunden. Das Suchen bestimmter Begriffe oder Namen (Zeichenkette) in einer Textdatei ist die einfachste Art des Information Retrieval.

Faktisch alle Textmanipulationsprogramme (Editoren) enthalten Kommandos (Funktionen, Optionen), mit denen der Benutzer Zeichenketten in der Datei suchen und sich auf dem Bildschirm oder Drucker ausgeben lassen kann. 'Zeichenkette' ist jede Folge von Zeichen, also auch mehrere Worte hintereinander oder auch nur ein einzelnes Zeichen.

Mit diesen Funktionen kann der Benutzer interaktiv einzelne Zeichenketten rückgewinnen. Oft ist aber die Suche nach nur einer Zeichenkette unbefriedigend, weil eben die Kooccurrenz, das gemeinsame Auftreten von Begriffen, einer der wichtigsten Analysepunkte in der Textverarbeitung ist. Textanalyseprogramme wie TEXTPACK V oder OCP enthalten Funktionen, die auch bedingten Information Retrieval erlauben. Allerdings sind diese Programme nicht interaktiv verwendbar und die Daten müssen vor dem Suchlauf für diese Programme aufbereitet sein. Schließlich muß der Benutzer noch die Kommandosprache des jeweiligen Systems lernen. Für kleinere Suchläufe oder kurzfristige Analysen bedeutet dies einen wesentlichen Zeitverzug.

## 2. SAR mit dem Siemens-EDT

Benutzer, die über ein programmiertes Textmanipulationsprogramm, wie z.B. Siemens-EDT, verfügen, können dieses auch für einfaches bedingtes Suchen verwenden. Die Bedingungen können der Form nach logische ODER- und UND-Bedingungen sein.

Der Einfachheit halber beschränke ich mich auf den Fall der UND-Verbindung. UND-Verbindung bedeutet, daß ein Textteil nur dann vom Programm ausgesucht wird, wenn in ihm zwei oder mehrere Zeichenketten zugleich vorhanden sind. Also z.B. alle Zeilen einer Zeitung, in der Schmidt und Genscher vorhanden sind.\*

Die Anweisung an das Programm lautet also:

1. nimm die erste Zeile des Textes und  
suche darin die Zeichenkette 'Schmidt'
2. falls dies nicht zutrifft: gehe zur nächsten Zeile; sonst:
3. suche in derselben Zeile 'Genscher'
4. falls dies nicht zutrifft: gehe zur nächsten Zeile, sonst:
5. gebe diese Zeile aus (Bildschirm, Drucker)
6. fahre fort, d.h. gehe wieder zu 1 und nimm dann die nächste Zeile.

---

\*Die Beschränkung auf Zeilen ist nicht notwendig, erfordert aber ein komplexeres Programm

Für den Siemens-EDT lautet eine solche

Anweisung:

```
Sp.1  
1. @ ON * FIND 'Schmidt'  
2. @ IF.FALSE. GO TO 6  
3. @ ON * FIND 'Genscher'  
4. @ IF.FALSE. GO TO 6  
5. @ PRINT *  
6. @ CONTINUE
```

Diese Anweisung wird in eine sogenannte 'Prozedur' des EDT geschrieben und ist nichts anderes als ein kleines Programm, mit dem die Textdatei manipuliert wird.

Um eine solche Prozedur anzulegen, gibt man im EDT das Kommando

```
@ PROC 1 (es gibt bis zu 21 Prozeduren für eine  
* @ ON *FIND. EDT-Sitzung, die mit PROC 1,  
: : PROC 2 ... angelegt werden können.  
: :  
@ CONTINUE '1' ist dann der 'Name' der Prozedur)  
@ END das Kommando ' END' schließt die Ein-  
gabe des Programms ab.**
```

---

NB. \*die doppelte Angabe von @@ ist notwendig, damit der EDT diese Zeile als Eingabe eines Textes und nicht als Kommando interpretiert.

\*\*eine Prozedur kann wie jede Textdatei geändert, erweitert etc. werden; vgl. dazu die einschlägigen Stellen im EDT-Handbuch.

Jetzt kann man, wenn man es noch nicht getan hat, die Textdatei in den Editor holen:

```
@GET 'TEXTDATEI' .
```

und mit

```
@DO 1 * = 1, $ , 1
```

das Suchprogramm starten.

Es gelten folgende Konventionen für das

@DO-Kommando:

\* = 1 ist die Angabe, daß die erste zu durchsuchende Zeile gleich der ersten Textzeile ist; \$ besagt, daß bis zur letzten Textzeile gesucht werden soll und die letzte '1' besagt, daß in ganzzahligen Sprüngen von Zeile 1 zu 2 zu 3 vorgegangen werden soll.\*

---

\* Der EDT kennt auch Zeilen mit Dezimalnummerierung, also Zeile 1.01, dann würde die Schrittweite nicht 1 sondern .01 betragen.

### 3. Erweiterungen der Prozedur

#### a) Drei und mehr Suchbegriffe

Sollen drei Zeichenketten berücksichtigt werden, so werden in der Prozedur zwei weitere Zeilen nach d) eingefügt, nämlich

```
4,1 @ON * FIND 'DRITTEN BEGRIFF'
```

```
4,2 @IF.FALSE. GO TO 8
```

und alle GO TO-Anweisungen lauten jetzt 8 (8. Zeile des Programms → CONTINUE; d.h. in 2., 4. und 4.2 muß '8' stehen.

#### b) Erleichterung des Auswechselns von Suchbegriffen

In einem weiteren Suchlauf, bei dem z.B. 'Genscher' gegen 'Brandt' ausgetauscht werden soll, muß der Benutzer mit

```
@PROC 1
```

wieder in die Prozedur 1 gehen und die Zeile 3 entsprechend ändern.

Dies kann er sich leichter machen, wenn er statt der expliziten Nennung der Begriffe Variablen einsetzt. EDT kennt 20 sogenannte 'Zeichenfolgevariablen' (~~#S1~~ - ~~#S20~~), in denen Zeichenketten abgelegt (gespeichert) werden können.

Die Zeilen 1. und 3. heißen dann

```
1. @ON * FIND #S1
```

```
2. @ON * FIND #S2
```

bevor die Prozedur mit

```
@DO 1 * = 1, # , 1
```

gestartet wird, muß der Benutzer angeben, für welche Zeichenketten ~~#S1~~ und ~~#S2~~ stehen sollen. Dies geschieht mit der Anweisung

```
@SET #S1 = 'Schmidt'  
@SET #S2 = 'Genscher'
```

beim nächsten Suchlauf, in dem 'Genscher' durch 'Brandt' ersetzt werden soll, heißt dann die Kommandoabfolge:

```
@SET S2 = 'Brandt'  
@DO 1 * = 1, # , 1
```

man braucht jetzt also nicht mehr in die Prozedur zurückzugehen.

c) Spaltenweises Suchen

Für Benutzer, die in bestimmten Spalten einer jeden Zeile bestimmte Informationen verschlüsselt haben (Fixed Format), kann die Suche auch auf Spalten innerhalb einer Zeile begrenzt werden.\*

Dazu muß die Angabe für die Zeile \* um die Angabe der jeweiligen Spalten ergänzt werden: \*: a-b  
a und b sind dann Ziffern und bezeichnen die Spalte ab der und bis zu der gesucht werden soll.

Hier ein Beispiel:

In der Studie 'Psychoanalytische Ambulanz' wurde für jede Zeile der Textdatei angegeben, welches Thema im Text abgehandelt wird. Die Angabe 'Thema' wurde als vierspaltiger Code in die Spalten 7-10 einer jeden Zeile geschrieben, der eigentliche Text steht in Spalte 12-80. Der Code für 'Diagnose' sei DIAG und es werden alle Diagnosen gesucht, in denen die Zeichenkette 'Psychoanalyse' vorkommen.

---

\* Diese Version ist auch besonders für numerische Dateien geeignet!

Dann lautet die Suchprozedur:

```
@@ON *: 7-10 FIND 'DIAG'  
@@IF.FALSE. GO TO 6  
@@ON *: 12-80 FIND 'PSYCHOANALYSE'  
@@IF.FALSE. GO TO 6  
@@ *  
@@CONTINUE
```

Durch die Beschränkung auf bestimmte Spaltenbereiche wird natürlich auch die Suchzeit verringert.

- d) Parameterübergabe (Schlüsselwörter) außerhalb der unter b. ist beschrieben, wie man mit dem SET-Kommando die Schlüsselwörter beim Ablauf der Prozedur eingeben kann. Der Vollständigkeit halber sei hier noch auf die Möglichkeit des EDT verwiesen, sogenannte Parameter, d.h. Leerstellen, in der Prozedur vorzusehen, die dann innerhalb eines erweiterten DO-Kommandos spezifiziert werden. Genaueres ist im EDT-Handbuch nachzulesen.

e) Sichern der Prozedur

Man kann die Prozeduren wie jede normale Datei mit einem SAVE-Kommando abspeichern und später mit einem GET-Kommando wieder laden:  
z.B.:

```
@ PROC 1  
.....  
....  
@SAVE 'PM.PROZEDUR.13'  
@END  
@PROC 14  
@GET 'PM.PROZEDUR.13'  
.....
```

Dann lautet die Suchprozedur:

```
@ON *: 7-10 FIND 'DIAG'  
@IF.FALSE. GO TO 6  
@ON *: 12-80 FIND 'PSYCHOANALYSE'  
@IF.FALSE. GO TO 6  
@*  
@CONTINUE
```

Durch die Beschränkung auf bestimmte Spaltenbereiche wird natürlich auch die Suchzeit verringert.

- d) Parameterübergabe (Schlüsselwörter) außerhalb der unter b. ist beschrieben, wie man mit dem SET-Kommando die Schlüsselwörter beim Ablauf der Prozedur eingeben kann. Der Vollständigkeit halber sei hier noch auf die Möglichkeit des EDT verwiesen, sogenannte Parameter, d.h. Leerstellen, in der Prozedur vorzusehen, die dann innerhalb eines erweiterten DO-Kommandos spezifiziert werden. Genaueres ist im EDT-Handbuch nachzulesen.

e) Sichern der Prozedur

Man kann die Prozeduren wie jede normale Datei mit einem SAVE-Kommando abspeichern und später mit einem GET-Kommando wieder laden:  
z.B.:

```
@PROC 1  
.....  
....  
@SAVE 'PM.PROZEDUR.13'  
@END  
@PROC 14  
@GET 'PM.PROZEDUR.13'  
.....
```

f) Protokollieren der Suchläufe

Die interaktiven Suchläufe können auf sehr einfache Weise protokolliert werden:

In BS 2000, dem Betriebssystem des Mannheimer Rechners, definiert man am besten zuerst eine Ausgabedatei mit dem Kommando

```
/SYSFILE SYSLIST = name der ausgabedatei
```

dann weist man das Betriebssystem an, alle Ein- und Ausgaben am Bildschirm in die 'ausgabedatei' zu protokollieren:

```
/OPTION MSG=FH
```

ruft man jetzt den Editor auf, so werden u.a. alle erfolgreichen Suchvorgänge in dieser Datei protokolliert.

Ist man mit der Suche fertig und hat das Programm EDT abgeschlossen (Kommando  $\odot$ H), so wird die Protokolldatei mit folgenden Kommandos ausgedruckt:

```
/OPTION MSG=F          - keine Protokollierung mehr  
                        der Ein-Ausgabe
```

```
/SYSFILE SYSLST = (PRIMARY)  
                  Rückschalten auf System-  
                  Druckfile
```

```
/PR   ausgabedatei, T = STATIONZ, LINES = 60*
```

---

\* T = STATIONZ - Ausgabe bei ZUMA  
 LINES = 60 - maximal 60 Zeilen pro Seite,  
 wegen der "Schönheit" des  
 Druckbildes.

NB. Alle /- Kommandos sind von Änderungen des Betriebssystems abhängig. Stand Juli 1982!

TEXTPACK V

Kurzbeschreibung

P.Ph. Mohler/C. Zül1

ZUMA  
Sept. 1983

## Einleitung

=====

TEXTPACK V ist ein Paket, das sowohl fuer die quantitativ ausgerichtete Inhaltsanalyse als auch fuer das Datenmanagement innerhalb qualitativer Textanalysen geeignet ist. Der groesste Teil der Programme dient der Erschliessung und Aufbereitung von Texten, nur zwei Programme (TAGCODER und SUBSEL) sind auf rein quantitatives Vorgehen ausgerichtet. Innerhalb des Bereichs der computerunterstuetzten Inhaltsanalyse (cui) ist TEXTPACK V sowohl fuer die Analyse mit dem reinen Woerterbuchansatz (a priori Kategorien, inhaltsanalytischer Diktionaer) als auch mit dem empirischer Ansatz (z.B. Cluster der haefigsten Woerter) geeignet.

Die neuen Optionen erlauben die Eingabe beliebiger Zeichen im Textteil, also auch Gross/Kleinschreibung, die Verarbeitung von Stammwoertern (d.h. erdungslose Woerter) und Mehrwortverbindungen mit den Programmen KWIC und TAGCODER (Beschreibung siehe unten), die getrennte Auszaehlung von Worthaefigkeiten von Untereinheiten eines Textes in einem Arbeitsschritt, die Begrenzung der Ausgabedateien durch GO-STOP-Wortlisten und andere Parameter sowie eine Reihe von Optionen zur Leberpruefung der automatischen Vercodung.

Das Paket umfasst derzeit folgende Programme:

1. SENTENCE - dieses Programm ueberfuehrt Rohtexte in eine festformatierte TEXTPACK-Systemdatei. Der Text kann mittels vom Benutzer festzulegender Sonderzeichen (z.B. dem Punkt ".") automatisch in Untereinheiten zerlegt werden (d.h. mit dem "." in Saetze als Untereinheiten).
2. SENMERGE - dies Programm unterstuetzt die Korrektur und Ergaenzung bestehender SENTENCE-Systemdateien.
3. LISTSPLT - zerlegt den Fliesstext einer SENTENCE-Datei in eine Datei von Einzelwoertern (Vertikaltext) und druckt gegebenenfalls den Text aus (zwei Druckformate).
4. FREQ - zaehlt Worthaefigkeiten ueber den gesamten Text bzw. dessen Untereinheiten aus und berechnet der Typ-Token-Ratio (TTR). Die Datei der Worthaefigkeiten kann durch verschiedene Parameter in ihrem Umfang eingeschaenkt werden.
5. FRELIST - dient zum Ausdrucken des mit FREQ erzeugten Haefigkeitsverzeichnisses in zwei Druckformaten und erzeugt wahlweise eine fuer den interaktiven Aufbau eines inhaltsanalytischen Diktionaers vorbereitete Datei der Worthaefigkeiten.
6. KWIC - mit diesem Programm koennen einzelne Woerter, Stammwoerter oder Mehrwortverbindungen in ihrem Kontext in der Form Schluesselwort in Kontext (KWIC) oder Schluesselwort ausserhalb des Kontextes (KWOC) dargestellt werden.

7. XREF - erzeugt ein Verzeichnis der Woerter eines Textes mit Verweisen auf die Fundstellen (Index). Der Umfang des Indexes kann mit wehrerer Optionen begrenzt werden.
8. WORDCOMP - mit diesem Programm kann der Wortschatz zweier Texte direkt verglichen werden.
9. TAGCODER - dient der Vercodung eines Textes mit Hilfe eines vorgegebenen Woerterverzeichnis (inhaltsanalytisches Diktionaer oder Liste der haeufigsten Woerter ect.). Das Programm erzeugt sowohl Dateien mit den Kategoriehaeufigkeiten als auch mit den Kategorieabfolgen. Die Validitaet der Vercodung kann mit mehreren Optionen, wie z.B. dem Einfuegen der Kategoriennummern in den laufenden Text, ueberprueft werden.
10. SUBSEL - stellt die Verbindung von numerischen Dateien, die Informationen ueber oder zu einer Textdatei enthalten, und der betreffenden Textdatei her. Aufgrund der Informationen in der numerischen Datei koennen die entsprechenden Texteinheiten ausgewaehlt werden. Das Programm kann zur Zeit SPSS-, OSIRIS- und Rohdateien verarbeiten.
11. REFORM - WANDELT EINE SENTENCE-DATEI IN CAPD IMAGE FORMAT fuer die Weitergabe bzw. Eingabe in andere Programme um.

#### Erlaeuterungen zur Begrifflichkeit

-----

In lockerer Reihenfolge erlaeuern wir hier die Begriffe, die wir in diesem Handbuch verwenden. Wir unterstellen dabei, dass der Benutzer die Klassiker der computerunterstuetzten Inhaltsanalyse kennt.

Zum leichteren Auffinden geben wir zuerst einen alphabetischen ueberblick ueber die einzelnen Begriffe:

Datei  
 Diktionaer  
 Disambiguierung  
 Ein- und Ausgabedateien  
 Ein- und Ausgabesaetze  
 Fließ- und Vertikaltext  
 Filter IDn=m  
 GO- STOP-listen  
 Identifikatoren  
 Kategorie  
 Monitor  
 Sortieren  
 Textaufbereitung  
 Textmanipulation mit TEXTPACK  
 Woerter  
 Zaehleinheiten  
 Zeichen  
 Zeichenketten

## Textaufbereitung

---

Die Textaufbereitung in der quantitativen Inhaltsanalyse unterscheidet sich nur im Ziel, nicht aber in der Technik von der in der qualitativen Textanalyse, die sich des Hilfsmittels Computer bedienen will. In beiden Faellen muessen zuerst die Texte in EDV-Dateien (im folgenden Textdatei) umgeformt werden (s. Datei). Ueblicherweise werden die Texte dazu in bestimmten Formaten auf computergerechte Medien (Datentraeger) uebertragen. Dazu gehoeren Lochkarter, Klarschriftleserformulare (OCR), direkte Eingabe am Bildschirm (direkt data entry). Schon getippte oder gedruckte Texte koennen mit der KDE<sup>m</sup> (Kurzweil)-Maschine direkt, ohne neues Abschreiben, eingegeben werden. Die Uebertragung von Texten ist niemals vcellig fehlerfrei, deshalb erfolgt als zweiter Schritt die Korrektur der Textdatei. Fuer jeden Computer gibt es mindestens ein Textaufbereitungsprogramm (engl. editing, Editor) wie z.B. WORDSTAR fuer viele Mikrocomputer, XEDIT fuer IBM, EDT und EDOR fuer Siemens. Diese Programme sind vielseitige und sehr bequem zu bedienende Instrumente. TEXTPACK enthaelt wegen der grossen Verbreitung dieser Programme und deren hohem Gebrauchswert nur ein rudimentaeres Editorprogramm (SEMERGE), mit dem neue Texteinheiten eingeflegt, alte gegen neue ausgetauscht oder alte geloescht werden koennen. Fuer alle anderen Formen der Textaufbereitung, von der Uebertragung in eine Datei bis zur Feinkorrektur, verweisen wir auf die jeweils vorhandener Editorprogramme. Nebenbei sei bemerkt, dass in gewisser Mass das TEXTPACK-Programm FPEQ die Textkorrektur erleichtert. Mit diesem Programm kann man ein Woerterverzeichnis (Vokabular) einer Textdatei erzeugen. Falsch geschriebene Woerter lassen sich nach unserer Erfahrung leichter beim Durchlesen eines Woerterverzeichnisses feststellen als beim Durchlesen des Textes selber. Das Ergebnis kann eine Liste von falsch geschriebenen Woertern sein, die man dann mit dem Editorprogramm korrigiert.

## Zeicher

---

An einigen Stellen der Programmbeschreibung sprechen wir von "Zeichen" und "Zeichenketten". Diese etwas schwerfaellige Sprache haengt mit der unterschiedlichen Behandlung von Ziffern (0-9), Buchstaben und Sonderzeichen (!?" usw.) im Computer zusammen.

In der guten alten Zeit des Computers gab es eigentlich nur Ziffern, denn es sollte ja gerechnet werden. Die Manipulation von Buchstaben und Sonderzeichen ist also erst ein spaetes Kind der Computer. Deshalb heissen Computer im guten Buerokratendeutsch heute auch nicht mehr Rechner, sondern elektronische Datenverarbeitungsanlagen, weil sie eben ausser rechnen noch einiges andere koennen. Die Verspaetung, mit der nun die Buchstaben in die Computer Einzug nahmen, hat fuer die Textverarbeitung bzw. Textanalyse einschneidende Folgen gehabt: fuer viele Sprachen, wie z.B. dem Deutschen, gibt es keine Normen fuer diejenigen Buchstaben, die nicht im englischen Alphabet vertreten sind (deshalb

ist dieses Manual auch "umlautlos" geschrieben). Neben diesem fundamentalen Problem gibt es fuer Handbuecher wie das unsere eines der Nomerklatur. Ziffern werden im forthinein auch als numerische Informationen bezeichnet. Die Unterscheidung zwischen Ziffern und dem allgemeinen Begriff Zeichen, der Ziffern, Buchstaben und Sonderzeichen umfasst, ist in TEXTPACK nur fuer die Textidentifikatoren (s. dort) von Bedeutung. Da mit Ziffern leichter gerechnet werden kann als mit Buchstaben und da gueltige Ziffern eindeutig nur die Zeichen 0-9 umfassen, werden formale Pruefungen, wie ununterbrochene, aufsteigende Reihenfolge der Sequenznummern, die Pruefung auf "gueltige" Zeichen (beides im Programm SENTENCE) und das Herausfiltern bestimmter Texteinheiten (Option IDN=m in fast allen Programmen) nur fuer Identifikatoren in Zifferndarstellung durchgefuehrt.

#### Zeichenketten

-----

Eine Zeichenkette ist eine beliebige Aneinanderreihung von Zeichen und darf in TEXTPACK zwar laenger als 39 Zeichen lang sein, nur werden alle Zeichen nach dem 39. Zeichen ignoriert. D.h., dass z.B. nicht zwischen den beiden woertern Conaudampfschiffartskapitaenslizenzaussteller und Conaudampfschiffartskapitaenslizenzausarbeiter unterschieden werden kann. Normalerweise bedeutet dies nur eine Beschraenkung fuer Mehrwortverbindungen, nicht aber fuer Woerter und Wortstaemme.

#### Woerter

-----

Bis jetzt ist immer von Woertern oder Wortformen gesprochen worden. Dies soll nun spezifiziert werden. Ein Wort wird in TEXTPACK als eine Folge von Zeichen definiert, die links und rechts von einem oder mehreren Leerzeichen begrenzt wird. Dem Leerzeichen oder Blank kommt also in TEXTPACK die Wortabgrenzungsfunktion zu. In der Beschreibung des Programms SENTENCE findet man Hinweise auf die Transkriptionsregeln von Texten, die sich aus dieser Definition ergeben. Ausser KWIC und TAGCODER operieren alle TEXTPACK-Programme mit dieser Wortdefinition. Diese beiden fuer viele Analysen zentralen Programme erlauben eine erweiterte Definitionen von woertern, naemlich wortstaemme und Mehrwortverbindungen:

ir TEXTPACK ist Wortstamm als "linker Teil eines Wortes" im Sinne der allgemeinen Wortdefinition definiert, wobei der linke Teil beliebig lang oder kurz sein kann (mindestens ein Zeichen). Beispiel: Der Wortstamm "arbeite" ist linker Teil der Woerter arbeiten, arbeitete, arbeitest. Der Stamm "arbeit" ist dazu noch linker Teil von Woertern wie arbeits- teilig, arbeitslos usw.

Mehrwortverbindungen sind definiert als eine unveraenderliche Aneinanderreihung einzelner Woerter (nicht von Wortstaermen!). Typische Beispiele sind Namen wie Jakob Fugger, Vereinigte Staaten oder feste Redewendungen wie freiheit-

liche Grundordnung, law and order.

Bei Mehrwortverbindungen, die mehr als 39 Zeichen enthalten, werden ebenso wie bei Wörtern, nur die ersten 39 Zeichen von TEXTPACK bearbeitet (vgl. Zeichen und Zeichenketten).

### Disambiguierung

-----

Das Programm TAGCODER encodet Wörter, Wortstämme und Mehrwortverbindungen grundsätzlich kontextfrei. So wird das Wort "Steuer" immer derselben Kategorie zugewiesen gleichgültig, ob "jemand das Steuer herumreisst" oder "jemand seine Steuer zahlt". Falls es fuer die Analyse von Bedeutung ist, dass nur die finanzielle Steuer encodet werden darf, so müssen die unterschiedlichen Bedeutungen der Zeichenkette "Steuer" vor dem Durchgang durch TAGCODER vom Benutzer disambiguiert, d.h. eindeutig gemacht werden, indem er z.B. Nummern (Steuer#1, Steuer#2) an die entsprechenden Wörter anfügt.

### Textmanipulationen mit TEXTPACK

-----

Nach der Feinkorrektur können die eigentlichen Textmanipulationen mit TEXTPACK beginnen. Im wesentlichen sind dies sechs Manipulationen: Wörterverzeichnis mit Worthäufigkeiten, Index, Schlüsselwörter im Kontext (Konkordanz), Vokabularvergleiche, Textselektion mit Hilfe textexterner Informationen und schließlich die Encodung. In den nächsten Abschnitten werden diese Manipulationen näher erläutert:

- a. Wörterverzeichnis mit Worthäufigkeiten  
ist ein einfaches Mittel der Textbeschreibung und bedeutet, dass fuer einen oder mehrere Texte die Häufigkeiten der verschiedenen Wortformen ausgezählt werden (Programme FREQ und FRELIST).
- b. Index  
ist ein Verzeichnis der Fundstellen von Wörtern in einem Text (Programm XREF).
- c. Schlüsselwörter im Kontext  
die Bedeutung einzelner Wörter ist oft nicht eindeutig. Deshalb ist es fuer Bedeutungsanalysen notwendig, Wörter in ihrem Umfeld (Kontext) aufzufinden. Das Programm KWIC (Key-Word-In-Context) erzeugt kontextbezogene Textauszüge fuer bestimmte Wörter, Wortstämme und Mehrwortverbindungen.
- d. Vokabularvergleich  
den Vergleich zweier Vokabulare auf gemeinsame bzw. unterschiedliche Begrifflichkeit erlaubt das Programm WORDCOMP (word comparison).
- e. Textselektion mit Hilfe textexterner Informationen  
hat man zu einem Text noch andere Informationen in Form

einer numerischen Datei (z.B. SPSS oder OSIRIS-Datei), dann kann man mit dem Programm SUBSEL Teile des Textes auswählen, die zu einer bestimmten Merkmalskombination in der numerischen Datei gehören.

#### f. Vercodung

schliesslich kann man festhalten, welche Wörter, Wortstämme und Mehrwortverbindungen in welchem Textteil vorkommen, dies als numerische Datei abspeichern und damit dann gegebenenfalls statistische Manipulationen mit SPSS oder OSIRIS durchföhren (Programm TAGCODER). In der Inhaltsanalyse wird dies gewöhnlich nicht fuer einzelne Wörter gemacht, sondern fuer Gruppen von Wörtern, die einen gemeinsamen Sinn teilen (inhaltsanalytische Kategorien).

#### Identifikatoren (IDs)

Unter "Textaufbereitung" wurden die "bestimmten computergerechten Formate" erwähnt, in die die Originaltexte ueberföhrt werden müssen, bevor sie mit Computern bearbeitet werden können. Das Prinzip ist bei allen Textprogrammen gleich. Einzelne Textteile müssen durch Kennzeichnung voneinander unterscheidbar sein. Solche Textteile können die einzelnen Wörter, Zeilen, Abschnitte, Kapitel usw. einer Textdatei sein. Identifikatoren sind nun Zeichenketten, die Informationen ueber bestimmte Textteile enthalten, z.B. Zeilennummern, Kapitelueberschriften oder ähnliches. Diese Identifikatoren enthalten also Informationen ueber den eigentlichen Text und werden von den Computerprogrammen fuer die Suche bzw. Kennzeichnung bestimmter Textteile benutzt. Weil sie nicht zum eigentlichen Text gehören, müssen sie fuer die Programme von diesem unterscheidbar sein. In TEXTPACK geschieht dies durch die räumliche Trennung von Identifikatoren und Text, d.h., sie stehen links und/oder rechts vom Text, quasi als Marginalien. Pro Arbeitsgang kann TEXTPACK nicht mehr als drei verschiedene Dimensionen der Identifikation bearbeiten (z.E. Dimension 1 = Buchtitel, Dimension 2 = Kapitel, Dimension 3 = Satz im Kapitel). Die Dimensionen sind entsprechend dem Beispiel immer hierarchisch anzuordnen. Dimension 1 ist also immer umfassender als Dimension 2 und diese als Dimension 3. Mindestens eine Dimension muss der Benutzer bei der Textaufbereitung angeben. Weitere Regeln und Restriktionen findet man in der Beschreibung des Programms SENTENCE.

Eine gewisse Verwirrung scheint fuer einige Benutzer die Beschränkung auf drei Dimensionen zu sein, denn oft gibt es mehr Dimensionen (z.B. Erscheinungsjahr, Buch, Kapitel, Absatz, Satz, Zeile, Wort). Die Beschränkung sagt nur, dass in einem Arbeitsgang maximal drei davon beruecksichtigt werden können, also einmal Jahr, Buch, Kapitel und zum zweiten Buch, Absatz, Satz. Jeder Schritt dabei bedeutet, dass eine neue SENTENCE-Datei zu erzeugen ist! Abgesehen davon, dass die Erzeugung so vieler Identifikatoren immer auf Kosten der Zeilenlänge des Textteils geht und recht mühsam ist, so ist die Neugenerierung einer neuen SENTENCE-Datei fuer jeden Analyseschritt ein ermüdendes

Geschaefit. Deshalb sei hier auf die Moeglichkeit des Programms SUBSEL verwiesen, mit dessen Hilfe aus einer bestehenden SENTENCE-Datei eine Textuntermenge aufgrund von Informationen aus einer numerischer Datei ausgewaehlt werden kann. Im obigen Beispiel waere es sinnvoll, Buch, Kapitel und Satz als Identifikatoren neber den Text zu schreiben:

A. BUCH KAPITEL SATZ TEXT-TEXT-TEXT

und eine zweite numerische Datei anzulegen, die die zusaetzlichen Informationen Jahr und Absatz enthaelt:

B. BUCH KAPITEL SATZ JAHR ABSATZ

Mit SUBSEL koennen dann z.B. aus allen Buechern des "Jahres" 1948 die ersten "Absaeetze" der dritten "Kapitel" ausgewaehlt werden. (Fuer dieses Auswaehlen benoetigt man Datentransformationsroutinen wie z.B. das SELECT IF von SFSS.)

#### Zaehleinheiten

-----

Unter dem Stichwort "Identifikatoren" wurde die bis zu drei Stufen tiefe Gliederung eines Textes in TEXTPACK erlaeuert (z.B. Buch- Kapitel- Satz). Bei der Vercodung mit dem Programm TAGCCDEP erhalten die Dimensionen der Gliederung eine zusaetzliche Bedeutung: sie werden zu potentiellen Zaehleinheiten. Man kann also das Buch als Zaehleinheit nehmen und zaehlen, wie oft die Kategorien des Diktionaers im Buch auftauchen. Nimmt man das Kapitel als Zaehleinheit oder den Satz, dann zaehlt das Programm dementsprechend kapitel- bzw. satzweise.

#### Filter IDn=#

-----

Gelegentlich will man nur bestimmte Textteile wie z.B. ein bestimmtes Buch oder nur die Textteile eines Interviews, die vom Befragten stammen, bearbeiten. Ist die notwendige Information in den Identifikatoren erhalten, dann kann man mit der Option IDn=# die entsprechenden Textteile direkt auswahlen (ist die Information in einer anderen Datei gespeichert, so kann man mit Hilfe von SUBSEL die entsprechenden Teile auswahlen). Angenommen, folgende Informationen seien in den Identifikatoren (IDs) enthalten: Interviewnummer (ID1) und Interviewer (ID2=1) bzw. Befragter (ID2=2), dann gibt man, um die Aeusserungen der Befragten auswahlen zu lassen, auf der Parameterkarte "OPTION ID2=2,...." an. Sollen darueber hinaus nur die ersten zehn Interviews bearbeitet werden, so lautet die Anweisung:  
OPTION ID1=1-10, ID2=2,....

#### Fliess- und Vertikaltext

-----

Dies ist ein Fliesstext und

dieses  
ist  
ein

#### Vertikaltext

Pro Zeile (Satz, Record) ein Wort abzuspeichern macht es dem Programmierer leichter (und die Programme effizienter), den Text zu manipulieren. Deshalb wird auch in TEXTPACK zumeist mit einem Vertikaltext (Ausgabedatei von LISTSPRT) gearbeitet.

#### Kategorie

Die Begriffe Wortliste und Kategorie werden in diesem Handbuch synonym gebraucht. Technisch gesehen handelt es sich dabei immer um eine Reihe von Wörtern, die wegen irgend eines gemeinsamen Merkmals zu einer Einheit zusammengefasst werden (z.B. "der, die, das" oder "gehen, laufen, rennen, schreiten"). In TEXTPACK haben sie die Form einer Liste:

```
001  der
001  die
001  das
017  gehen
017  laufen
017  rennen
```

Die ersten drei Ziffern bezeichnen dann die Nummer einer Kategorie. Mehrere Kategorien, die fuer eine Studie entwickelt werden, bilden dann das Messinstrument, den inhaltsanalytischen Diktionaer oder das inhaltsanalytische Klassifikationsschema. Der Begriff "Diktionaer" verweist auf den technischen Aspekt, dass diese Woerterlisten eine aehnliche Struktur wie ein Synonymwoerterbuch (Thesaurus) haben und ist eine direkte Uebersetzung des englischen "dictionary". Der Begriff "Klassifikationsschema" verweist eher auf den methodologischen Status der Woerterlisten, naemlich einzelne Woerter, Wortstaemme oder Mehrwortverbindungen zu klassifizieren.

#### Diktionaer

Diktionaer ist die Bezeichnung fuer ein Klassifikationsschema in der computerunterstuetzten Inhaltsanalyse. Dieses Schema besteht aus einzelnen Listen von Woertern (Kategorien) und wird wegen seiner Aehnlichkeit mit Synonymwoerterbuechern Diktionaer genannt (s. Kategorie).

#### GO-STOP-Listen

In fast allen Programmen TEXTPACKs gibt es die Option "GO" oder "STOP" und der Verweis auf entsprechende Wortlisten. Dahinter verbirgt sich ein sehr einfaches Prinzip der Datenreduktion. Oft will man nicht jedes Wort einer Textdatei in die Textmanipulation miteinbeziehen. Ein Beispiel da-

fuer sind die Satzzeichen: d.h., Satzzeichen muessen nach den im Programm SENTENCE angegebenen Schreibregeln vom Wort abgetrennt und zwischen zwei Leerzeichen gesetzt werden. Damit werden Satzzeichen nach der Wortdefinition TEXTPACKS zu selbststaendigen Woertern. Nur mag es ja anfaenglich ganz interessant sein, wieviele Punkte, Kommas, Klammern, Gedankenstriche usw. es in der Textdatei gibt. Doch, abgesehen von der Computerzeit, die es kostet, bei jeder Haeufigkeitszaehlung auch noch jedesmal die Satzzeichen mitauszuzaehlen, ist es einfach ermuedend, sich jedesmal aus den ueberfluessigen Informationen die wichtigen herauszusuchen. Im Falle der Satzzeichen gibt man deshalb eine Liste aller Zeichen, die man unbeachtet lassen will (STOP-Liste). Man kann diese Liste natuerlich auch um sogenannte Funktionsworte (z.B. Artikel, Konjunktionen) oder jedes andere Wort, das fuer die jeweilige Analyse nicht interessiert, erweitern.

Andererseits will man oft nur bestimmte Woerter analysieren (z.B. ihre Haeufigkeiten in Textuntereinheiten betrachten). Dann gibt man eben eine Liste dieser Woerter an und erreicht durch die Option GO, dass nur diese und keine anderen manipuliert werden sollen. Formal gesehen ist eine GO- oder STOP-Liste nichts anderes als eine spezifische Form eines Klassifikationsschemas oder Diktionaers.

#### Sortieren

-----

Das Sortieren der Textdatei ist ein aeusserst wichtiger Teil der Textmanipulationen mit TEXTPACK. Computertechnisch ist sortieren nicht trivial. Je nach Maschinentyp kann ein bestimmtes Sortierverfahren sehr schnell oder extrem langsam, manchmal gar unmoeglich sein. Deshalb hat TEXTPACK kein eigenes Sortierprogramm, sondern hier wird der Benutzer aehnlich wie bei der Textaufbereitung (Editoren) auf das jeweilige allgemeine Sortierprogramm seines Computers verwiesen (bei Siemens heisst es z.B. SORT).

Diese Programme arbeiten nun alle nach demselben Prinzip: der Benutzer muss sogenannte Sortierfelder und die Sortierfolge angeben. Um die Begriffe Sortierfelder und Sortierfolge anschaulicher zu machen, nehmen wir als Beispiel eine Mitgliederdatei, die aus folgender Elemente besteht:

1. Mitgliedsnummer
2. Nachname
3. Vorname

Die Elemente stehen alle auf einer Zeile und immer an selben Platz:

```
001 Mueller Hannah
002 Mayer Karl
003 Schulze Friedrich
004 Mueller Johann
```

d.h., die Mitgliedsnummer beginnt in der ersten Position (Spalte bei einer Lochkarte) und geht bis zur dritten Position, der Nachname steht in der 5. bis 14. Position (d.h., er kann maximal 10 Zeichen lang sein) und der Vorname steht in der 15. bis 24. Position.

Mitgliedsnummer, Nachname und Vorname belegen drei Sortier-

felder in der Datei, die jetzt nach der Mitgliedsnummer aufsteigend (von 1 bis 4) sortiert sind. Eine Umsortierung nach absteigender Sortierfolge ergaebe die Reihenfolge 004,003,002,001.

Soll die Datei jetzt nach den Nachnamen alphabetisch aufsteigend sortiert werden, so muss man dem Sortierprogramm mitteilen, dass das Sortierfeld in Position 5 beginnt, bis zu 14 Zeichen lang ist und die Sortierfolge aufsteigend sein soll. Fuer den Siemens SORT heisst das z.B.:

```
/EXEC $SORT
  SORT FIELDS=(5,14,A)
  END
```

Das Ergebnis waere die Reihenfolge Mayer, Mueller, Mueller, Schulze.

Soll ueber Nachnamen und Vornamen sortiert werden, also mehrere Sortierfelder beim Sortieren beruecksichtigt werden, dann legt die Reihenfolge der Sortierfelder die endgueltige Sortierung fest; gibt man VORNAME, NACHNAME an, so ergibt sich:

```
003 Schulze      Friedrich
001 Mueller      Hannah
004 Mueller      Johann
002 Mayer        Karl
```

Dreht man die Reihenfolge um, also NACHNAME, VORNAME, so ergibt sich:

```
002 Mayer        Karl
001 Mueller      Hannah
004 Mueller      Johann
003 Schulze      Friedrich
```

Die letzte Angabe waerde fuer den SIEMENS SORT folgendermassen lauten:

```
/EXEC $SORT
  SORT FIELDS=(5,10,A,15,10,A)
  END
```

Im TEXTPACK muss die Reihenfolge der Sortierfelder insbesondere bei den Programmen FREQ, KWIC und TAGCODER beachtet werden, um absurde Ergebnisse zu verhindern (z.B. Worthaeufigkeiten fuer jeden Satz eines Buches!). Fuer gewoehnlich sind die Identifikatoren (ID1, ID2, ID3) und das Einzelwort die Sortierfelder in TEXTPACK. Gelegentlich koennen auch die Kategoriennummern oder die Position eines Wortes innerhalb einer ID-Einheit Sortierfelder sein. Fast immer ist die Sortierfolge aufsteigend.

Die Positionen der Sortierfelder findet man in der Tabelle im Anhang oder in den jeweiligen Programmbeschreibungen.

#### Datei -----

Eine Datei (engl. file) bezeichnet in der Computerwelt eine systematisch angeordnete Menge zusammengehoeiger Informationen. Frueher wurden Dateien in Form von Lochkarten oder Lochstreifen ausserhalb des Computers aufbewahrt, waren also

im wahrsten Sinne des Wortes greifbar. Zur Zeit sind Magnetplatten und Magnetbänder die bevorzugten Speichermedien. Für den TEXTPACK-Benutzer ist es nicht so wichtig, in welcher Form seine Dateien gespeichert sind (ausser dass 20000 Lochkarten äusserst schwer zu tragen sind), sondern welche Art Dateien er für TEXTPACK benötigt. Er braucht mindestens eine Textdatei, in der der zu analysierende Text gespeichert ist. Diese Datei kann er mit den verschiedenen Programmen manipulieren. Einige Programme wie KWIC oder TAGCODER benötigen eine zweite Datei, in der einzelne Textwörter gespeichert sind, die im Text gesucht werden sollen (Diktionaer).

GO-STOP-Lister sind ebenfalls Dateien, die Wörter enthalten, die, wenn im Text gefunden, besonders behandelt werden sollen. Die Programme erzeugen z.B. Dateien der Wörter der Textdatei zusammen mit ihren Häufigkeiten (FREQ), bestimmter Wörter im Kontext (KWIC), den Textuntereinheiten (SUBSEL), der Codes (TAGCODER) und ähnliches mehr. Der Benutzer muss für jedes Programm, sofern er keinen Monitor zur Verfügung hat, der ihm die Arbeit abnimmt, angeben, wie die Ein- und Ausgabedateien heissen.

#### Ein- und Ausgabedateien

-----

Die TEXTPACK-Programme sind eigentlich nur eine Art Fleischwolf. Oben kommt ein ganzes Stück hinein und unten das Gehackte raus. Und wie im Fleischwolf wird etwas hineingegeben und etwas ausgegeben: "Eingabe" und "Ausgabe" (Input, Output). Im technischen Sinne sind beides Dateien, z.B. eine SENTENCE-Datei, die in das Programm LISTSPLT hineingesteckt wird und als SPLIT-Datei herauskommt. Damit das Programm weiss, was Eingabe ist und wohin die Ausgabe gesteckt werden soll, teilt man ihm dies mit einem besonderen Kommando mit: Bei IBM ist es das DD-Kommando, bei SIEMENS das FILE-Kommando, bei UNIVAC das ASC-Kommando usw.

Immer wird dabei eine Datei mit dem Eingang oder Ausgang eines TEXTPACK-Programms verbunden. Ein- und Ausgabedateien werden immer durch eine 2-stellige Zahl, der Nummer der Ein- und Ausgabereinheit gekennzeichnet. SENTENCE z.B. schreibt die Ausgabe für gewöhnlich auf die Einheit 20. Diese Nummern können in allen Programmen durch entsprechende Optionen ("INFILE", "OUTFILE") geändert werden.

#### Ein-Ausgabesätze

-----

Im TEXTPACK-Manual wird der Begriff "Satz" im doppelten Sinn verwendet. Der erste ist der "Textsatz" im leblichen Sinne (engl. sentence) und bedarf keiner weiteren Erläuterung. Der zweite bezeichnet die kleinste Einheit einer Ein- und Ausgabedatei (engl. record) und entspricht einer einzelnen Lochkarte oder einer Textzeile. In einigen Computern, wie dem SIEMENS-Computer wird zwischen Sätzen (records) fester Länge und variabel langen Sätzen unterschieden. Sätze fester Länge sind wie Lochkarten, die 80 Spalten haben,

immer gleich lang (d.h., sie werden mit Leerstellen aufgefüllt), variabel lange Sätze sind wie die Zeilen eines handgeschriebenen Textes: ausgefranst. Ob die eine oder andere Satzart genommen wird, hat technische Gründe und kann vom Benutzer nicht veraendert werden. Nicht-SIEMENS-Benutzer sollten die besonderen Hinweise im Anhang studieren.

#### Monitor

Ein Monitor ist so etwas wie ein freundlicher Fuehrer durch den Computer. Er ist ein besonderes Programm, das dem Benutzer die Arbeit erleichtert, indem es z.E. die Dateien (Ein-/Ausgabe) verwaltet und das Sortieren uebernimmt. Leider ist es sehr aufwendig oder gar unmoeglich, einen guten Monitor fuer alle Computertypen zu bauen und deshalb haben wir uns vorerst darauf beschaenkt, einen Monitor fuer unsere Maschine, eine SIEMENS, zu programmieren. Im Anhang, der die computerspezifischer Hinweise enthaelt, ist vermerkt, ob es fuer der Computer, den Sie benutzen, einen Monitor gibt, der dann von einem Spezialisten fuer ihre Maschine geschrieben wurde.

## SENTENCE

=====

## Allgemeine Beschreibung

-----

Das Programm erstellt eine Datei im TEXTPACK-Systemformat. Ein solcher Systemfile ist Voraussetzung fuer die weitere Verarbeitung von Texten mit TEXTPACK. Das Programm erwartet maschinenlesbare Textdaten als Eingabe. Die Gliederung des Texts in verschiedene Zaehleinheiten muss durch die Vergabe von Identifikatoren angezeigt werden.

## Moeglichkeiten des Programms

-----

Das Programm verarbeitet maschinenlesbare Texte, die auf Karte, Band oder Platte gespeichert sein koennen. Das jeweilige Speichermedium wird durch einen entsprechenden Parameter angegeben (INPLT).

Die Identifikatoren legen die Zaehleinheiten fest. Die Identifikatoren koennen aus beliebigen Zeichenketten bestehen. In der Regel sollten jedoch Ziffern, also Identifikationsnummern verwendet werden, da alphanumerische Identifikatoren den Gebrauch der ID-Filteroptionen ausschliessen! Die Textdatenbasis wird durch die Identifikatoren hierarchisch gegliedert. TEXTPACK erlaubt bis zu drei Ebenen (im folgenden Zaehleinheit genannt) der Gliederung (ID1, ID2, ID3). ID1 muss immer angegeben werden, die Untergliederung in ID2 und ID3 ist optional. Die Identifikatoren fuer die Ebenen ID1 und ID2 koennen maximal sechs Zeichen umfassen, ID3 kann bis zu fuef Zeichen lang sein. Alle numerischen Identifikatoren muessen jeweils rechtsbuerdig angegeben werden.

Der Text innerhalb der jeweils kleinsten Zaehleinheit kann durch fortlaufende Zeilennummern (maximal 5 stellig) gekennzeichnet (SEQ) sein.

Die Optionen PHRASE und MARKER ermoeglichen die automatische Generierung fortlaufender numerischer Identifikatoren fuer die Zaehleinheiten ID2 oder ID3.

Die Positionen der Identifikatoren, der Sequenznummern und des Texts innerhalb eines Eingabesatzes werden mit Parametern angegeben (DIM1, DIM2, DIM3, SEQ, TEXT). Ein Eingabesatz darf nicht laenger als 256 Zeichen sein (diese Angabe gilt fuer die Siemens Version, Beschraenkungen fuer andere Rechner sind installationsspezifisch).

SENTENCE kann mehrere Eingabedateien (die alle das gleiche Format haben) verarbeiten und zu einer Ausgabefile verbinden.

Schon beim Einlesen der Daten koennen bestimmte Textteile durch Filteroptionen auf der Basis der Identifikatoren ausgewählt werden.

SENTENCE komprimiert nicht die einzelnen Texte durch Eliminieren mehrfacher Leerzeichen (Blanks) innerhalb des Textes. Entfernt werden nur Leerstellen am Anfang einer Textzeile und am Ende der Zeile. Es ist Sache des Benutzers, zu entscheiden, ob diese mehrfachen Blanks zu Fehlern bei

der Analyse fuehren koennen (vgl. KWIC, TAGCODER, TEXTCODE) und gegebenenfalls diese mit Hilfe eines Editors zu beseitigen.

Die praktische Erfahrung zeigt, dass Eingabedateien in der Regel Fehler aufweisen. Diese Fehler koennen in zwei Gruppen eingeteilt werden:

1. in formale Fehler, die im Identifikator-Teil der Eingabedaten auftreten, und
2. in Transkriptionsfehler, die beim Uebertragen der Texte in ein maschinenlesbares Format entstehen.

SENTENCE identifiziert nur formale Fehler und diese nur bei Zifferndarstellung (numerisch) der Identifikatoren. Je nach Fehlertyp werden vier unterschiedliche Fehlermeldungen ausgedruckt. Die Fehlermeldungen werden im Abschnitt "Druckerausgabe" naeher beschrieben. Treten formale Fehler auf, so wird die vollstaendige Information ueber die fehlerhafte Einheit (Identifikationsnummer und Text) ab der fehlerhaften Stelle bis zum Ende ausgedruckt. Die entsprechende Fehlerkorrektur kann mit dem Programm SENMERGE (vgl. Programm SENMERGE) vorgenommen werden.

## SENMERGE =====

### Allgemeine Beschreibung -----

Mit SENMERGE koennen Zaehleinheiten (=kleinste ID-Ebene) in SENTENCE-Dateien ausgetauscht, ergaenzt oder geloescht werden.

### Moeglichkeiten des Programms -----

SENMERGE korrigiert und ergaenzt bestimmte Zaehleinheiten, ohne dass die gesamten Texteingabedaten neu eingelesen werden muessen.

Es koennen erstens fehlerhafte Zaehleinheiten korrigiert (ausgetauscht), zweitens fehlende Zaehleinheiten hinzugefuegt und drittens vorhandene Zaehleinheiten geloescht werden.

Die Korrektur-Datei (zweite Eingabedatei) kann in SENTENCE-Format oder als Roh-Textdatei vorliegen. Die Eingabe einer Roh-Korrekturdatei ist allerdings nur bei wenigen Korrekturen sinnvoll. Ansonsten empfiehlt es sich in der Regel, die Korrekturen mit dem Programm SENTENCE zu bearbeiten und danach via SENMERGE mit der Hauptdatei zu verknuepfen, da die formale Fehlerpruefung in SENTENCE umfangreicher ist als in SENMERGE.

REFORM  
=====

### Allgemeine Beschreibung

-----

REFORM (REFORMAT system file) setzt eine SENTENCE-Datei in eine Datei um, deren Satzlaenge vom Benutzer festgelegt werden kann. Damit koerner SENTENCE-Dateien leicht vor einem Programm zum anderen (z.B. COCOA, OCP) transportiert werden.

### Moeglichkeiten des Programms

-----

Der Benutzer kann mit dem Parameter LENGTH eine beliebige Satzlaenge zwischen 60 und 256 festlegen. Da die Identifikatoren in einer SENTENCE-Datei immer 6 (ID1, ID2) bzw. 5 Zeichen lang sind, oft aber die tatsaechlich in der Datei auftretenden Identifikatoren kuertzer sind, koennen diese mit den Parametern CUT1, CUT2, CUT3 verkuerzt ausgegeben werden. Es werden nur soviele Identifikationsnummern in die Ausgabe uebernommen, wie tatsaechlich in der Datei gespeichert sind oder vom Benutzer verlangt werden.

Bei der Eingabe in SENTENCE gilt die Trennungsregel, dass Zeichenketten, die mit einem Bindestrich enden, in der naechsten Zeile fortgesetzt werden. Falls der Benutzer eine dementsprechende Ausgabe wuenscht, so ist der Parameter WRAP anzugeben, d.h., es wird mitten im Wort abgetrochen und das Textfeld in der Zeile mit einem Bindestrich als letztem Zeichen abgeschlossen. Das Wort wird in der naechsten Zeile fortgesetzt.

LISTSPLT  
=====

## Allgemeine Beschreibung

-----

Das Programm LISTSPLT druckt Textdateien aus und/oder erzeugt aus einer Fliesstextdatei (Ausgabe des Programms SENTENCE) eine Vertikaltextdatei. Diese Vertikaltextdatei dient als Eingabe in die Programme FREG, XHEFF, KWIC, WORDCOMP und TACCCDES.

## Moeglichkeiten des Programms

-----

Mit der Option LIST koennen SENTENCE-Dateien ganz oder teilweise (globaler Filter ID=nn) ueber den Schnellrucker im Format DIN 44 oder DIN 17 ausgegeben werden.

Mit der Option SPLIT werden SENTENCE-Dateien, die der Text in Form eines Fliesstextes enthalten, in Einzelwortdateien (Vertikaltext) zerlegt. Ein logischer Satz (Record) einer SPLIT-Datei enthaelt dann nur noch ein Wort und seine Herkunftsverweise (siehe Ausgabe).

Die SPLIT-Datei kann durch die explizite Angabe von Identifikationsnummern (ID=nn) begrenzt werden (globaler Filter). Eine weitere Moeglichkeit, die SPLIT-Datei zu begrenzen, ist die genaue Angabe von GO- bzw. STOP-Woertern. Danach enthaelt die SPLIT-Datei genau die Woerter, die in der GO-Liste stehen bzw. genau die Woerter nicht, die auf der STOP-Liste stehen (lokaler Filter).

Es ist moeglich, eine durch eine GO-/STOP-Liste verkleinerte SPLIT-Datei als FOR-Datei in SENTENCE einzulesen und standardgemaess zu bearbeiten. So vorzugehen empfiehlt sich besonders bei sehr grossen Dateien, die durch die Eliminierung besonders haeufiger und fuer viele Analysen entbehrlicher Woerter (z.B. Artikel, Konjunktionen) erheblich reduziert werden koennen.

Besonders zu beachten ist, dass Sonder- und Satzzeichen in Version V immer als Einzelwoerter behandelt werden (siehe Fulleitung). Mit Hilfe der STOP-Liste kann ihre Verarbeitung unterdrueckt werden. Dies entspricht der Option P\*ARK in Version IV.

FREQ  
====

### Allgemeine Beschreibung

-----

Mit FREQ werden die unterschiedlichen Woerter einer Textdatei ausgezaehlt (Erstellen eines Woerterverzeichnis). Zusaetzlich kann der Type-Token-Ratio (d.h. das Verhaeltnis unterschiedlicher Woerter zur Gesamtzahl aller Woerter eines Textes) berechnet werden.

### Moeglichkeiten des Programms

-----

Es koennen Woerterverzeichnisse der gesamten Datei oder einzelner Untereinheiten (je nach Vorgabe auf der Ebene von ID1 oder ID2 oder ID3) erstellt werden.

Nicht immer ist ein Verzeichnis **a l l e r** Woerter (Types) sinnvoll. Deshalb kann das Verzeichnis in FREQ durch STOP- oder GO-Listen und/oder die Vorgabe minimaler bzw. maximaler Laefigkeiten begrenzt werden. Eine STOP-Liste enthaelt alle Woerter, die nicht in das Woerterverzeichnis aufgenommen werden sollen. GO-Listen hingegen enthalten nur die Woerter, die aufgenommen werden sollen.

Oftmals ist es notwendig, nicht nur das Woerterverzeichnis des gesamten Korpus, sondern auch das einzelner Untereinheiten zu kennen. Mit der Option SEPARATE koennen solche Unterverzeichnisse erzeugt werden. Diese Verzeichnisse koennen auf allen drei Ebenen der Identifikationsnummer erstellt werden, allerdings mit der Einschraenkung, dass pro Lauf nur eine Ebene bearbeitet werden kann.

Eine der gebraeuchlichsten Masszahlen fuer die Komplexitaet von Texten ist das Verhaeltnis von unterschiedlichen Wortformen zur Gesamtzahl der Woerter eines Textes (Type-Token-Ratio, TTR). FREQ berechnet diese Masszahl und gibt sie ueber den Drucker aus. Wird das Woerterverzeichnis durch die Parameter GO, STOP, MAXI oder MINI begrenzt, so wird dennoch der TTR bezogen auf alle Woerter ausgegeben.

FRELIST  
 =====

### Allgemeine Beschreibung

-----

Mit dem Programm koennen zum einen die in einer Textdatei vorkommenden woerter, Zahlen und Sonderzeichen (im folgenden "Woerter" genannt) zusammen mit der Haeufigkeit ihres Vorkommens gelistet werden und zum anderen eine Datei dieser Wortformen erzeugt werden, die so formatiert ist, dass sie fuer die interaktive Entwicklung eines inhaltsanalytischen Diktionaers verwendet werden kann.

Eingabe fuer FRELIST ist die Ausgabe von FREQ (siehe Programm FREQ).

### Moeglichkeiten des Programms

-----

Soll das Woerterverzeichnis einer Textdatei gedruckt werden (Liste der Wortformen (types) zusammen mit ihren jeweiligen Haeufigkeiten), so koerner mit FRELIST folgende Drucktilder erzeugt werden:

- a. Standardmaessig werden 2 woerter pro Zeile gedruckt.
- b. mit der Option FCRMAT=DINA4 wird 1 wort pro Zeile (in DIN A4 Format) gedruckt.

Fuer die Konstruktion eines cbi-Diktionaers kann das mit FREQ erzeugte woerterverzeichnis in einem vom Benutzer festgelegten format als latei ausgegeben werden.

XREFF  
=====

### Allgemeine Beschreibung

-----

Das Programm XREFF (Crossreferences) erzeugt ein Verzeichnis aller Wörter einer Textdatei (Index). Für jedes Wort (Type) werden die Identifikationsnummern und die Position des Worts innerhalb der niedrigsten Identifikationsnummer angegeben. Die Ausgabe von XREFF ist eine Druckliste, Eingabe ist ein alphabetisch sortierter Vertikaltext (Ausgabe von LISTSPLT).

### Möglichkeiten des Programms

-----

Oft ist es sinnvoll neben Häufigkeitslisten von Wörtern (siehe FREQ) oder KWIC-Listen (siehe KWIC) auch einen Index aller oder eines Teils der Wörter anzulegen, um so im Fliesstext (siehe LISTSPLT) bestimmte Wörter in ihrem vollen Kontext lokalisieren zu können. Die Ausgabe von XREFF kann durch die Parameter IDN=(n) und GC/STOP begrenzt werden. Sonderzeichen und Satzzeichen werden wie Wörter behandelt. Um die Druckerausgabe einzuschränken, empfiehlt es sich in der Regel, Satzzeichen, Artikel usw. durch eine STOP-Liste auszuschliessen.

KWIC  
 ===

### Allgemeine Beschreibung

-----

Mit dem Programm KWIC (Key Word In Context) koennen einzelne Woerter oder Zeichenketten (im folgenden Schluesselwoerter) in ihrer jeweiligen Kontext dargestellt werden (Konkordanz). Der Kontextumfang kann durch Parameter gesteuert werden. Die Schluesselwoerter werden in Form eines TAGCODER-Diktionaers eingegeben.

Eingabe in KWIC ist entweder je eine SENTENCE- und SPLIT-Datei oder eine KWIC-Zeilen-Datei.

Ausgabe ist eine Druckliste und wahlweise eine festformatierte Datei von KWIC-Zeilen.

### Moeglichkeiten des Programms

-----

Das Programm arbeitet mit zwei unterschiedlichen Kontextdefinitionen. Die erste definiert als Kontext die kleinste Identifikationsebene und innerhalb dieser Ebene bis zu 45 Zeichen vor und nach dem Schluesselwort. Demnach ist die Druckerausgabe immer eine Zeile pro Fundstelle und die Schluesselwoerter beginnen immer ab Spalte 71.

Die zweite definiert als Kontext wiederum die kleinste Identifikatorebene und innerhalb dieser Ebene diejenige Einheit, in der das Schluesselwort steht. Dieser Kontext kann um die vorangehende und die folgende Einheit erweitert werden (siehe Option UNIT). Das Schluesselwort wird bei der zweiten Definition nicht im Kontext hervorgehoben. Stattdessen steht es in der Druckerausgabe als Ueberschrift ueber allen seinen Fundstellen, ist also ausserhalb des Kontextes hervorgehoben. Deshalb nennt man die zweite Kontextdefinition Key Word Out of Context (KWOC) im Gegensatz zu Key Word In Context (KWIC) der ersten Definition.

Die Schluesselwoerter werden im Format des TAGCODER-Diktionaers eingegeben (siehe Eingabe). Es koennen ausser Einzelwoerter auch Zeichenketten, die nicht laenger als 39 Zeichen lang sind, angegeben werden. Die Zeichenketten muessen durch ein "+" oder "-" in der Spalte 5 gekennzeichnet sein (siehe Eingabe). Die Formatgleichheit mit dem TAGCODER-Diktionaer ermoeglicht, die Gueltigkeit eines inhaltsanalytischen Diktionaers auf einfache Weise zu ueberpruefen.

Sollen Wortverbindungen wie z.B. 'John Smith' gesucht werden, so ist zu beachten, dass Leerstellen (Blanks) als Zeichen gezahlt werden. Da TEXTPACK die Eingabe nicht auf Leerstellen ueberprueft, hat der Benutzer selbst sicher zu stellen, dass zwischen Wortverbindungen immer gleich viele Blanks sind (vergleiche Programm SENTENCE, Eingabe).

Mit der Option AMBIG koennen mehrfach vercodete Woerter bzw. Zeichenketten ueberprueft werden. Mehrfach vercodet heisst, dass ein Wort gleichzeitig mehreren Kategorien eines Diktionaers angehoert.

Laeuft das Programm unter der Option KWIC, so kann eine fest formatierte Datei ausgegeben werden, die nach unterschiedli-

chen Kriterien sortiert (z.B. Identifikatoren, Kategorien, Schlüsselwoertern, Kontext links oder rechts) und in einem zweiten Durchgang durch KWIC ausgedruckt werden kann (siehe Option INKWIC).

Hinweise:

- a) Die Option KWOC kann zu uebermaessig langen Ausgabelisten fuehren.
- b) Die Kategorien eines TAGCODER- Diktionaers koennen mit der Option HUNT im Programm TEXTCODE aehnlich der Option KWOC gesucht und gelistet werden.

**WORDCOMP**  
=====**Allgemeine Beschreibung**  
-----

Das Programm WORDCOMP vergleicht den Wortschatz zweier Textdateien. Es berechnet die absoluten Differenzen der Wort-häufigkeiten zwischen den Texten.

**Möglichkeiten des Programms**  
-----

Unterscheidet sich das Vokabular zweier Texte, enthält ein neuer Text Wörter, die noch nicht im Klassifikationschema (Diktionaer) enthalten sind? WORDCOMP unterstützt die Beantwortung solcher Forschungsfragen. Die Vokabulare von zwei Texten werden verglichen und die Differenzen der Wort-häufigkeiten zusammen mit einer Wortliste ausgegeben.

Das Programm hat folgende Optionen:

1. Vollständige Auflistung aller Wörter (Types) beider Texte, deren Häufigkeiten und der entsprechenden Differenzen (Option ALL);
2. Auflistung der Wörter, die nur in einem der beiden Texte enthalten sind (Option NEW);
3. Begrenzung aller Angaben durch GO/STOP Wortlisten (Option CC/STOP);
4. Auswahl nach Identifikationsnummern.

SUBSEL  
=====

## Allgemeine Beschreibung

-----

SUBSEL stellt die Verbindung zwischen Textdaten und anderen -zumeist numerischen- Daten her. Mit dem Programm koennen aus den Textdaten Untereinheiten ausgewaehlt werden. Die Auswahl erfolgt anhand der Auspraegungen einer Variablen (Kriteriumsvariable) der numerischen Datei. Die Textdatei muss im SENTENCE-Format vorliegen, die numerische Datei kann eine OSIRIS III, eine SPSS 9 oder eine "rechteckige" Rohdaten-Datei sein (vgl. Programm SENTENCE und die Manuals zu SPSS und OSIRIS III, fuer andere Systeme ist die dortige Rohdatenausgabe oder das jeweilige SPSS/OSIRIS Interface zu aktivieren).

## Moeglichkeiten des Programms

-----

Das Programm SUBSEL prueft in der numerischen Datei, ob die von Benutzer angegebene Kriteriumsvariable den oder die von ihr festgesetzten Werte hat. Trifft diese zu, so werden alle Saetze (Records) der Textdatei, die dieselben Identifikatoren (IDs) wie der entsprechende Satz der numerischen Datei haben, ausgewaehlt. Die Text-Records werden im SENTENCE-Format ausgegeben. Zusaetzlich kann der Benutzer noch ueber den ueftlichen Filter (ICN=\*) bestimmte Textteile von der Bearbeitung ausschliessen.

SUBSEL hat keine eigenen dem RECODE anderer Programme aehnlichen Optionen. Liegt eine Kriteriumsvariable nicht schon in einer Form vor, die eine bequeme Selektion ermoeglicht (z.B. dichotomisierte Variablen: Hochschulabschluss ja/nein), so muss bevor SUBSEL eingesetzt werden kann, mit den Datentransformationsoptionen des jeweiligen Programmsystems, in dem die numerischen Daten vorliegen, eine bestehende Variable recodiert oder eine neue erzeugt werden. Fuer die Auswahl mit Hilfe der Kriteriumsvariablen ist die Angabe eines oder mehrerer Werte bzw. Wertebereiche zulaessig (z.B. "1" oder "1-12" oder "1,5,12" oder "1-4,6,88-90"). Waehrend der TEXTPACK-Filter (ICN=\*) dazu dienen kann, einzelne Faelle oder Fallgruppen, deren Identifikatoren bekannt sind auszuwaehlen, dient SUBSEL dazu, Faelle aus der Textdatei auszuwaehlen, deren Identifikatoren im Vorhinein nicht bekannt sind. So zum Beispiel die Antworten von Befragten, die auf einer Skala bestimmte Werte erreichen ("Postmaterialisten", "Liberale", etc.) oder Texte, in denen eine bestimmte Kategorienkorstellation besonders haeufig vorkommt.

## Auswahl der Texteinheiten

-----

Der Benutzer gibt eine Kriteriumsvariable an, deren Werte zur Auswahl fuehren sollen. Das Programm prueft dann jeden Satz der numerischen Datei, ob dort die Kriteriumsvariable den oder die angegebenen Wert(e) hat. Trifft dies zu, so

sucht das Programm in der Textdatei diejenige Texteinheit, die dieselben Identifikatoren wie der Fall der numerischen Datei hat. Die Identifikatoren der Texteinheit sind die IDs (ID1, ID2, ID3), fuer die numerische Datei muss der Benutzer angeben, welche Variablen als Identifikatoren dienen sollen. Es muss mindestens der der ID1 entsprechende Identifikator angegeben werden.

Besonders wichtig ist das Aggregationsniveau der numerischen Datei. Um logische Fehler zu vermeiden, sollte folgender Grundsatz beruecksichtigt werden:

Das Aggregationsniveau der numerischen Datei ist gleich oder groesser als das der Textdatei.

Dazu ein Beispiel:

Die Jubilaeumsausgabe Goethes in sechs Baenden sei als Textdatei vorhanden. Die Textdatei ist untergliedert in 1. Baeude (ID1), 2. Stuecke (ID2) und 3. Kapitel in den Stuecken (ID3). Zu der Textdatei gibt es eine numerische Datei, die Informationen, wie "Erscheinungsjahr", "Zahl der Einzelworte", usw. enthaelt. Gibt es fuer jeden der sechs Baende nur einen summarischen Record in der numerischen Datei, dann gibt SUBSEL pro Treffer in der Kriteriumsvariablen den entsprechenden Fund als Texteinheit aus. Gibt es zu jeder Stueck einen numerischen Record, so werden dementsprechend Stuecke und Kapitel ausgegeben. Liegt nun pro Textsatz eines Kapitels noch ein numerischer Record vor, dann gibt es in der numerischen Datei mehrere Records mit den gleichen Identifikatoren fuer ID1, ID2 und ID3. Letzteres bedeutet, dass das Aggregationsniveau der numerischen Datei kleiner als das der Textdatei ist. Bei der Ausfuehrung stoesst SUBSEL dann auf mehrere Records der numerischen Datei, die dieselben Identifikatoren haben. Das Programm beruecksichtigt dann nur den ersten dieser Records und druckt eine Warnung aus. Dies kann zu irrtuemlichen Fehlern fuehren, wenn zum Beispiel alle Kapitel gesucht werden, in denen die Kategorie "Teufel" auftritt. Da der erste Satz im Faust diese Kategorie nicht enthaelt, wuerde demnach der erste Akt (=Kapitel) des Faust nicht beruecksichtigt werden. Eine weitere Fehlermoeglichkeit ergibt sich, wenn der Benutzer eine beliebige, nicht ID1, ID2 oder ID3 entsprechende Variable faelschlicherweise als Identifikator angibt. In diesem Fall ergeben sich Zufallstreffer, die formal nicht als Fehler zu erkennen sind.

## TAGCODER

=====

## Allgemeine Beschreibung

Das Programm TAGCODER verschlüsselt Texte an Hand eines inhaltsanalytischen Diktionsaers. Der Diktionsaer besteht aus Listen von Woertern, Wortfolger oder Zeichenketten, jede Liste ist die Operationalisierung einer inhaltsanalytischen Kategorie. Zaehleinheiten werden durch die Identifikatoren (ID) im Text und die Wahl einer der drei Identifikatorebenen (ID1, ID2 oder ID3) festgelegt. Das Programm hat zwei Vercodungsmodi (TAE, VEC siehe unten). Eingabe in TAGCODER ist eine sortierte SPLIT-Datei (siehe Sortierung). Moegliche Ausgaben sind fest formatierte Dateien der numerischen Codes, eine Datei der richtvercodeten Zaehleinheiten, eine Datei in der die jeweils zutreffenden Kategoriennummern in den Text eingefuegt sind bzw. die entsprechenden Woerter ersetzen und ein Setup fuer die Weiterverarbeitung der numerischer Dateien mit SPSS, OSIRIS oder SIR.

## Moeglichkeiten des Programms

## 1. Prinzip der Vercodung

Das Programm vergleicht die Elemente (tags) jeweils einer Kategorie des Diktionsaers mit einer Zaehleinheit des Textes. Die Zaehleinheit wird mit dem Parameter CODEGR festgelegt und kann je nach Spezifikation eine der drei ID-Dimensionen (ID1, ID2, ID3) umfassen. Die Kategorie ist in der Regel eine Woerterliste (z.B. ich, du, er, sie, es). Statt einzelner Woerter koennen auch Zeichenketten (feste Wortverbindungen, z.P. Hans Schmidt, bzw. Teile von Woertern, z.E. polit) Elemente einer Liste sein. Der Verknuepfungsoperator zwischen den Elementen einer Kategorie ist das logische ODER. Wenn ein Element einer Kategorie in der Zaehleinheit angetroffen wird, fuehrt dies zur Vercodung. Tritt ein und dasselbe Element mehrmals in einer Zaehleinheit auf oder erscheinen verschiedene Elemente einer Kategorie in einer Zaehleinheit, so wird jedes Auftreten vercodet (gezahlt). Elemente koennen mehreren Kategorien angehoren (sogenannte mehrdeutige Woerter). Wird ein solches Element im Text gefunden, so wird ein Code fuer jede einzelne seiner Kategorien vergeben. Die Zaehleinheiten werden durch die Identifikatoren (ID1, ID2, ID3) begrenzt. Der Benutzer kann festlegen, auf welcher Ebene der Identifikatoren ausgezahlt werden soll (CODEGR=IDn). Beispiel: ID1 bezeichnet Bucher, ID2 Kapitel und ID3 Saetze. Bei CODEGR=ID3 wird dann pro Satz gezahlt bzw. vercodet; bei CODEGR=ID2 pro Kapitel und bei CODEGR=ID1 pro Buch. Die Zaehlebene ist je nach Analysestrategie festzulegen.

## 2. Vercodungsmodi

Das Programm kann in zwei Richtungen vercoden. Einmal zaehlt es pro Zaehleinheit die Haeufigkeit von Kategorien aus (TAB). Die numerische Ausgabedatei enthaelt dann die Haeufigkeiten der Kategorien pro Zaehleinheit. Zum anderen vercodet das Programm die Abfolge von Kategorien (VEC). Dieser Vercodungsmodus ist mit der Mehrfachnennung in Umfragen vergleichbar. Dann enthaelt die Ausgabedatei Kategoriennummern in der Reihenfolge ihres Auftretens im Text.

Der Benutzer definiert fuer die Vercodung im TAB-Modus den Bereich moeglicher Kategoriennummern (z.B. 1-15 oder 1-999 oder 35-185, vgl. Option CATRANGE). Die groesstmoeegliche Zahl von Kategorien haengt von der Zahl der Stellen pro Kategorie ab (siehe Beschraenkungen).

Fuer die Vercodung im VEC-Modus gibt der Benutzer die groesstmoeegliche Zahl von Kategorienfolgen (Nennungen) an (z.B. 19 oder 999, vgl. Option MAXCODES). Gibt der Benutzer z.B. 5 als Begrenzung an, so schreibt das Programm in die Ausgabedatei maximal 5 Kategoriennummern pro Zaehleinheit in der Reihenfolge ihres Auftretens. Werden weniger als 5 Codes vergeben, werden die restlichen Positionen mit Nullen aufgefuellt. Auch hier ergibt sich die groesstmoeegliche Zahl der Nennungen aus der Feldweite, die fuer eine Kategoriennummer bereitgestellt wird (siehe Beschraenkungen).

## 3. Hilfsmittel zur Validierung der Vercodung

Die Gueltigkeit der Vercodung kann ausser mit den im folgenden aufgefuehrten Moeglichkeiten auch mit dem Programm KWIC ueberprueft werden.

### a. Option WRITEBACK

Mit dieser Option koennen die Kategoriennummern jeweils hinter (WRITEBACK=INSERT) das verschluesselte Element bzw. an dessen Stelle (WRITEBACK=REPLACE) in den Text geschrieben werden. Um die Kategoriennummer eindeutig von "gewoehnlichen" Textworten unterscheidbar zu machen, koennen bis zu zwei Zeichen als Praefix der Kategoriennummern festgelegt werden (z.B. PREFIX=x, dann wird aus Nummer 001 die Zeichenfolge x001). Der Benutzer kann die mit WRITEBACK erzeugte Ausgabedatei, die im SENTENCE-Format geschrieben wird, mit dem Programm LISTSPLIT (Option LIST) ausdrucken lassen.

### b. Option NOCODE

Fuer Zaehleinheiten, deren ueberhaupt kein Code zugewiesen werden kann, werden bei der Option NOCODE die Identifikationen ausgegeben. Mit dem Programm SLPSEL kann eine SENTENCE-Datei dieser Zaehleinheiten erzeugt und diese mit dem Programm LISTSPLIT ausgedruckt werden.

### c. Option HUNT

Gelegentlich ist es sinnvoll zu erfahren, welche Elemente einer Kategorie in einer Zaehleinheit vercodet werden. Mit HUNT erhaelt man eine Liste, in der pro Zaehleinheit die Elemente der Kategorien ausge-

ben werden. Die Option HLNT ist nur sinnvoll fuer die Analyse von einzelnen Zaehleinheiten. Groessere Teile oder der gesamte Text kcerren effizienter mit den Programmen KWIC und WORDCOMP im Sinne der HUNT-Option analysiert werden.



MÖGLICHKEITEN EDV-GESTÜTZTER STRUKTURIERUNG UND ORDNUNG  
QUALITATIVER INTERVIEWS

- Integrationsmöglichkeiten von Standardsoftware der  
Textverarbeitung mit kompatiblen FORTRAN-Programmen -

Vortragsskizze für ZUMA-Arbeitstagung (Qualshop) vom  
28.2. - 2.3.83

Januar 1983

Jürgen Schupp  
Tel.: 06155/4833  
Fichtestr. 6  
Griesheim

## GLIEDERUNG

	SEITE	
1.0	EINLEITUNG	1
2.0	EDV-GESTÜTZTES ARBEITEN MIT INTERVIEWTEXTEN	1
2.1	Abgrenzung des TMS zu anderen Textanalysepaketen	1
2.2	Texterfassung und -aufbereitung	2
2.2.1	Transkription	2
2.2.2	Eingabe	3
2.2.3	Kleinrechner	4
2.2.4	Nutzung von Standardsoftware der Textverarbeitung	6
2.2.5	Textkorrektur	8
2.2.6	Numerierung	8
2.2.7	Ausgabe	9
2.3	EDV-technische Probleme bei der Arbeit mit Texten	9
2.3.1	Transskriptionstexte aus EDV-technischer Sicht	9
2.3.2	Speicherplatz	11
2.3.3	Zugriffszeit	12
2.4	Programmtechnische Lösung	13
2.4.1	Aufgabe eines Textmanagementsystems (TMS)	14
2.5	Zum Umgang mit dem TMS	15
2.5.1	Auswertungspraktische Ziele des (TMS)	15
2.5.2	TOPICS	16
2.5.3	Summaries	20
2.5.4	weitere Programmfunktionen	21
3.0	ZUSAMMENFASSUNG DER VORTEILE VON EDV	22

### Literaturverzeichnis

#### Anlagen:

- 1: schematische Ablaufbeschreibung Phase I
- 2: schematische Ablaufbeschreibung Phase II
- 3a: Beispielblatt eines erfaßten Transkripts (Schnelldrucker)
- 3b: Beispielblatt eines erfaßten Transkripts (Typenraddrucker)
- 4: automatisierte Indizierung (Beispiel)

### Vorbemerkung

Der vorliegende Text wurde mit Hilfe eines Computers erstellt und editiert, wobei der Zeichensatz für deutsche Typenräder GERMAN COURIR 10 verwendet wurde, der auch Umlaute und ß kennt. Die editierte Fassung wurde mit Hilfe des auch im folgenden diskutierten Textverarbeitungsprogramms RUNDEU (dt. Fassung von RUNOFF, Digital Equipment Coop. (1975a) in der vorliegenden Form aufbereitet.

#### 1.0 EINLEITUNG

Die folgenden Abschnitte versuchen die Erfahrungen mit EDV-gestützter Textverarbeitung und -strukturierung vorzustellen, die im Rahmen einer Diplomarbeit (1) gesammelt wurden. Darüberhinaus sollen allgemeine Prinzipien von Textverarbeitung als Teil der Informationstechnik herausgearbeitet werden und die spezifischen Vorteile für eine Auswertung qualitativer Daten diskutiert werden.

#### 2.0 EDV-GESTÜTZTES ARBEITEN MIT INTERVIEWTEXTEN

Im folgenden werden Lösungen diskutiert, die am Hochschulrechenzentrum der Universität Frankfurt am Großrechner DEC 1091 praktisch umgesetzt wurden. Die Anwendungen sind von daher als systemspezifisch zu bezeichnen; dennoch zeigen die Erfahrungen an anderen Anlagen, daß eine Umsetzung bzw. Anwendung auf anderen Rechnertypen zumindest im Bereich von Universitätsrechenzentren mit umfangreichem standardmäßigem Software-Angebot nur wenig Aufwand bereiten würde.

#### 2.1 Abgrenzung zu anderen Textanalysepaketen

Der Computer diente bei den hier vorgestellten Anwendungen nicht als "General Inquirer" (2), der bestimmte

---

1) Schupp, J. (1982), es handelt sich um die Auswertung von 17 'Leitfadeninterviews' mit einer Gesamtgesprächszeit von ca. 24 Std.

Wortsequenzen aufspürt, statistische Auswertungen über die verwendeten Ausdrücke anfertigt, oder der die syntaktische Struktur der Sprache zu entschlüsseln sucht. Die hier vorgestellte Lösung steht also in keiner direkten Beziehung zu Verfahren der quantitativen automatisierten Inhaltsanalyse (3). Der Computer diene vielmehr in der Phase der Texterfassung zunächst rein als Arbeitsmittel - als flexible und ökonomische Alternative zur Schreibmaschine.

In der Strukturierungsphase der Interviews sollte ein vom Autor programmiertes Textmanagementsystem als Hilfsmittel zur ökonomischen Verwaltung von Texten auf der Basis von Interaktionen dienen. Es sollte eine flexible Handhabung eines Ordnungssystems für die inhaltliche Einteilung des Textes ermöglichen und darüber hinaus jederzeit eine Informationsrückgewinnung (auch im Kontext!) garantieren, also z.B. die Möglichkeit bieten, zu überprüfen, ob eine Äußerung des Befragten spontan erfolgte, oder ob sie die entsprechende Antwort auf eine spezifische Frage war.

## 2.2 Texterfassung und -aufbereitung

Nach der Feldphase wurden die Interviewtexte bis sie zu Auswertungszwecken zur Verfügung standen in mehreren Arbeitsschritten (vgl. schematische Darstellung Anhang 1) aufbereitet.

### 2.2.1 Transkription

Der erste Arbeitsschritt bildet die Erstellung von Transkripten. Entscheidet man sich zu einer EDV-Erfassung der Texte so sind mehrere Möglichkeiten einer Transkription denkbar, die von den jeweiligen personellen als auch ma-

---

2) vgl. Stone, J.Ph. u.a. (1966), S. 68 ff

3) EDV-spezifische Einzelheiten hierzu in:  
Kriz, J. 1975), Kap. Inhaltsanalyse, S. 147ff

teriellen Gegebenheiten abhängen.

- a) Der wohl aufwendigste Fall ist die handschriftliche Erstellung einer Rohtranskription jedes Interviews und die anschließende Eingabe der Texte am Bildschirm.
- b) Eine weniger aufwendige Methode setzt einen 'ruhigen' Arbeitsplatz an einem Terminal (4) voraus. Hier könnte über Kopfhörer direkt am Bildschirm transkribiert werden.
- c) Eine dritte Methode wäre die Erstellung von Transkripten mittels OCR-Kugelkopfschreibmaschinen, welche eine spätere Übertragung auf Datenträger erlauben.

### 2.2.2 Eingabe

Die Eingabe von Transkripten an einem Terminal (Fall a und b) setzt die rudimentäre Kenntnis eines EDITORS (5) des jeweiligen Betriebssystems der EDV-Anlage voraus.

Erfahrungsgemäß eignen sich hierzu sog. SCREEN-Editoren besonders gut, da hier die Vorteile der freien Cursorsteuerung (6) ausgenutzt werden können. Weiterhin sollte

- 
- 4) Terminals werden erfahrungsgemäß in allen Benutzern zugänglichen Terminalräumen aufgestellt, in denen mehrere Benutzer gleichzeitig arbeiten und daher mit gewissem Geräuschpegel zu rechnen ist.
  - 5) Zur Erstellung und Bearbeitung von Text-Dateien (also auch Transkriptionstexten) stellt jedes Betriebssystem zunächst einen EDITOR zur Verfügung, mit dessen Hilfe ein Benutzer am Terminal durch Anwendung einer Kommandosprache auch Textstellen innerhalb eines Interviews suchen, löschen oder verändern kann. Diese Editoren haben jedoch den Nachteil, dass sie betriebssystemspezifisch sind, und damit auch unterschiedlich schwierig oder komfortabel in der Handhabung.
  - 6) Damit man sich auf dem Bildschirm buchstaben- und zeilengenau orientieren kann, wird vom Betriebssystem automatisch ein Lichtpunkt (CURSOR) auf dem Terminal mitgeführt, der immer nach dem zuletzt geschriebenen Buchstaben erscheint.  
'Freie Cursorsteuerung' bedeutet: will man einen Text einrücken, im Wort korrigieren, oder löschen, lässt sich der Cursor an jede beliebige Stelle der 24 Zeilen mit je 80 Spalten 'fahren', um die Verbesserung vorzunehmen.

man Möglichkeiten der Zeilenschaltung benutzen, um die Eingabe der Texte ohne Rücksicht auf Zeilenende vornehmen zu können. Zudem bestehen bei den meisten Systemen Möglichkeiten des automatischen 'Backups' (7), um auch bei niemals auszuschließenden Systemabstürzen keine Arbeit doppelt verrichten zu müssen. Für regelmäßig wiederkehrende Zeichenkombinationen (Bsp. Interviewer: ) können bei vielen Editoren unbenutzte Zeichen des Terminaltastensfeldes 'umdefiniert' werden in die Wiedergabe eines solchen spezifischen Textsegments, bei dem Tippfehler dann auch ausgeschlossen sind.

Eine Eingabe von Texten, die mit OCR-Kugelkopfschreibmaschinen erstellt wurden, geschieht mittels sog. Klarschriftbelegleser; gewöhnlich mit Magnetband als Zwischenspeicher zum Großrechner (8).

Allen Eingabemöglichkeiten gemeinsam sollte das Resultat sein; nämlich die Verfügbarkeit von Magnetplattendateien mit symbolischem Inhalt; mit anderen Worten Transkriptionstexte auf Datenträger.

### 2.2.3 Kleinrechner

An dieser Stelle sollte man auch auf sog. Kleincomputer aufmerksam machen, die gewöhnlich Off-line (9) betrieben werden und als Speichermedium Disketten (floppy-disk) verwenden, worauf trotz stark eingeschränkter Speicherfähigkeit dennoch mindestens 1 komplettes 'qualitatives Interview' ab-

---

7) so kann man beispielsweise beim SED-Editor die Zahl der Characters angeben nach denen ein automatischer SAVE des Buffers auf Platte vorgenommen wird.  
(vgl. Hall, Ch. (1981), S. 45)

8) Details siehe Dworatschek, S. (1977)

9) Verarbeitungsart ohne direkte (aktive) Verbindung zur DV-Anlage.

speicherbar ist. Die Eingabe erfolgt auch hier (10) über komfortable Editoren. Der Vorteil des Off-line Betriebs entsteht durch die geringe 'Response Zeit' (11), allerdings für eine spätere programmgestützte Auswertung der Interviews ist eine Kopplungsmöglichkeit (Schnittstelle) an einen Großrechner nötig.

Existierende Großrechner sind meist auf Anforderungen des technisch-wissenschaftlichen Bereichs zugeschnitten; neue Benutzeranforderungen etwa auf dem Gebiet der Textverarbeitung wurden von ihnen oft nicht in zufriedenstellender Weise unterstützt, obgleich sich auch hier Verbesserungen durchsetzen. Auf der anderen Seite wurden ständig neue kleiner Rechner konzipiert, die spezielle Aufgaben (und nur diese) wesentlich kostengünstiger und meist auch schneller als eine universelle Rechenanlage abwickeln. Der Gedanke ist von daher naheliegend, mehrere spezialisierte Geräte zu koppeln (vernetzen) und somit die Gesamtheit aller Dienstleistungen der EDV zu nutzen.

Eine Kopplung für unser Anwendungsgebiet muß man sich derart vorstellen, daß die Möglichkeit bestehen muß, Interviewtranskriptionsdateien von einem Kleinrechner auf einen Großrechner zu 'kopieren' (12).

- 
- 10) bsp. beim MINC-11 von Digital Equipment Coop. ein Kleincomputer mit 2 Diskettenlaufwerken.
  - 11) bezeichnet die Zeit, die von einer Eingabe in den Computer bis zur Rückmeldung verstreicht. Bei TIME-SHARING Anlagen kann sich diese Zeit ab einer gewissen Anzahl von Teilnehmern sprunghaft erhöhen.
  - 12) Umfangreiche Netzwerksoftware der Rechnerhersteller erlaubt mittlerweile gerade bei editierten 'symbolischen' Dateien (also z.B. Texte) problemlosen Transfer.

#### 2.2.4 Nutzung von Standardsoftware der Textverarbeitung

Befindet sich der Transkriptionstext auf Datenträger, so wird die Struktur des Textes durch artifizielle Zeichen wie Zwischenraum ('blanks') oder Zeilenvorschub ('line') im Rechner bestimmt.

Um zur eigenlichen Struktur des Textes in Seiten, Absätzen, Zeilen zurückzukehren und um zu einer Ausgabe auf Papier zu gelangen, bedarf es lediglich der Kenntnis weniger Steuerbefehle (Print-Kommando).

Interessant erscheinen jedoch die Vorteile der Textverarbeitung mit Elementen des Textmanagements zu verbinden. Als Textverarbeitung wird hier die aus der Bürotechnik entstammende Bedeutung verwendet (13), die sich stark von Begriffsbestimmungen der Linguistik und Inhaltsanalyse abgrenzen.

Als Beispiel für eine solche Verwendung kann das Textverarbeitungsprogramm RUNOFF (14) genannt werden. Der Eingabetext wird mit einer Reihe von Steuerzeichen versehen, welche aus der Transkriptionsdatei einen druckfähigen Text

---

13) "Tätigkeit, die das Konzipieren, Formulieren, Diktieren, Schreiben, Reproduzieren, Transportieren und Archivieren sowie das Organisieren dieser Aufgabe umfasst", aus: Wossidlo, P.R. (Hrsg.), (1980), S. III, von Informatikern um die Elemente des Informationstransfers sowie der Zeichenkettenverarbeitung wesentlich erweitert.

14) Das Programm RUNOFF ist ein Programmsystem, das zur Aufbereitung und Ausgabe von Texten dient. Ausführliche Beschreibung in Digital Equipment Coop. (1975a). Das System ist auf anderen Computersystemen ebenfalls verfügbar. (so z.B. auch auf IBM 370-168 MVS/TSO)

entstehen lassen.

Außer verschiedenen Textaufbereitungsfunktionen wie Paragraphenbildung, Seitennumerierung, (Sub-)Kopfzeilen, sowie automatisierte Reservierung von Fußnoten kann insbesondere noch Randausgleich und Silbentrennung vorgenommen werden. Für unser Problem Transkriptionstexte sind zunächst vor allem die Paragraphenbildung sowie Titulierung interessant.

Die Eingabe der Steuerzeichen in den Text kann mit Zuhilfenahme des Editors erfolgen. Ein Parameterdefinitionsloop (PARDEF) (15) der im Prinzip mit jedem EDITOR definiert werden kann, durchläuft den Text und anschließend entsteht nach Aktivierung des Standardprogramms RUNOFF ein Ausgabe-text des gesamten Interviews, wie er in Anlage 3 auszugsweise beiliegt.

Es sei nochmals betont, daß eine Verwendung dieser Standardsoftware für die später vorgestellten Strukturierung des Interviewmaterials nicht zwingend ist; der Vorteil liegt neben dem besser 'ansprechenden' Schriftbild des Textes auch darin, daß bei der Texteingabe weniger Sorgfalt nötig ist.

Jedoch sei noch eine weitere Möglichkeit angesprochen, mit der sinnvolle Anwendungen von Textverarbeitungssystemen ohne jegliche Programmierarbeit erzielt werden könnten.

- 
- 15) Die Schleife PARDEF enthält folgende programmierte Editoranweisungen:
- I - Suche I: oder B:
  - II - gehe an Anfang der Zeile
  - III - Schiebe eine Leerzeile ein
  - IV - Schreibe in diese Zeile .P -10,1,2  
(Steuerzeichen für Paragraphenbildung)
  - V - gehe hinter :

Es handelt sich dabei um Indizierungen von Transkriptionstexten und die anschließende automatisierte Erstellung textbezogener 'Schlüsselwortlisten'. Die Idee wäre folgende: Bereits bei der Eingabe werden Begriffe bzw. Ausdrücke, die für eine spätere Interpretation relevant erscheinen könnten, zusätzlich mit einem sog. INDEX FLAG versehen (16), welcher bei der anschließenden Textverarbeitung eine alphabetisch sortierte Begriffsliste pro Interview liefert mit jeweiliger Seitennummer(n), wo der Begriff im Interview zu finden ist.

Bei Bedarf besteht zudem noch die Möglichkeit den Begriff fett gedruckt im Text hervorzuheben.

#### 2.2.5 Textkorrektur

Ein weiterer Arbeitsschritt bildete die Textkorrektur von Transkripten, um spätere Interpretationen nicht auf 'Mißverständnissen' aufzubauen.

Für die Erstellung einer 'Korrekturvorlage' des Textes genügt ein Ausdruck über Schnelldrucker, der ein Interview von 50 Seiten in weniger als 3 min. ausgibt. Im Prinzip könnte dieser Arbeitsschritt ganz am Bildschirm erfolgen; ohne Papier. Aus ergonomischen Gründen wird jedoch davor bei umfangreichen Texten abgeraten. Die Korrektur 'am' Text erfolgt dann wieder mit Hilfe des oben schon benutzten Editors, womit man durch 'Such-Anweisungen' schnell an die zu korrigierende Textstelle 'springen' kann.

#### 2.2.6 Numerierung

Nach Korrektur der Dateien wurden die Texte durch das FORTRAN-Programm NUMMER (Programmlogik siehe in Anhang 1) verareitet, wobei eine nach Aussagen durchnumerierte Version

---

16) z.B. > beim RUNOFF-System  
vgl. auch Anhang 4 mit einem Beispielausdruck

des Textes erstellt wurde. Da eine spätere Verwaltung der Texte auf Basis einzelner Interaktionen von Interviewer und Befragtem vorgesehen ist, leistet dieses Programm die hierfür notwendige Aufbereitung für die Abspeicherung im Textmanagementsystem (siehe Abschnitt 2.4.)

### 2.2.7 Ausgabe

Eine 'Reinschrift' jedes durchnummerierten Interviews erzeugte anschließend ein DIABLO 1640 Typenraddrucker. Dieser Druckertyp 'arbeitet' zwar wesentlich langsamer als ein Schnelldrucker. Das Schriftbild ist jedoch wesentlich deutlicher als bei Schnelldruckern (vgl. Anlage 3a und 3b). Technologische Fortschritte (Bsp. Laserdrucker) versprechen, daß in absehbarer Zeit auch an Hochschulrechenzentren 'schnell' druckfertige Texte erzeugbar sind.

## 2.3 EDV-technische Probleme bei der Arbeit mit Texten

Für eine weitere Verwendung der abgespeicherten Texte zu inhaltlichen Zwecken ergeben sich eine Reihe von grundsätzlichen Problemen, die hier kurz skizziert werden sollen.

### 2.3.1 Transskriptionstexte aus EDV-technischer Sicht

Nach Eingabe und Korrektur des Interviewtextes in einer symbolischen Datei, läßt sich deren Aufbau folgendermaßen beschreiben.

#### DATEIAUFBAU

- I)     .T INTERVIEW XYZ  
       .FLAG HYPHEN ...

```
II)      .P -10,1,2

III)     00001 I: text
           text
           text

II)      .P -10,1,2

III)     00002 B: text
           text
           text
           text

II)      .P -10,1,2

III)     00003 I: text
           text

           u.s.w.
```

Die Datei beginnt zunächst mit Steuerzeichen (I) für das Textverarbeitungssystem RUNOFF, die das Format, Überschrift und Silbentrennzeichen spezifizieren. Zwischen den Interaktionen von Interviewer und Befragtem stehen wiederum Steuerzeichen (II), welche die Interaktion beim Druck optisch voneinander absetzen (Unterbrechen von Zeilenfüllung).

Jede Interaktion (III) beginnt mit einer 5-stelligen Zahl in der 1. Spalte einer Zeile. Sie ist erst dann beendet, wenn in der 1. Spalte erneut ein Steuerzeichen folgt, was die Interaktion von der folgenden abgrenzt. In der EDV-Terminologie nennt man eine solche Einheit logischen Satz. Das Kennzeichen der hier abgespeicherten logischen Sätze ist ihre variable Länge, das heißt, der Text benötigt unterschiedlich viel Speicherplatz, wenn er sequentiell hintereinandersteht. Für die im folgenden entwickelte

Lösung stellt eine Interaktion die kleinste Einheit dar, die vom Textmanagementsystem zugreifbar ist. Die Entscheidung darüber was eine Zählereinheit darstellt, oder man könnte auch anlehnend an die Terminologie der quantitativen Inhaltsanalyse 'Analyseeinheit' sagen (18), mußte bereits vor der Texterfassung getroffen werden. Praktisch sah diese Entscheidung so aus, daß die Zeichenfolge "I: " bzw. "B: " beginnend in der 1. Spalte einer Zeile dem Programm NUMMER die Anweisung gab, einen neuen logischen Satz (d.h. eine neue 5-stellige Nummer) zu definieren, der obiges Bild (III) besitzt.

Eine Abgrenzung nach kleineren Einheiten (syntaktische Sätze oder nur Halbsätze) ist programmtechnisch genauso leicht umsetzbar, muß jedoch, um es deutlich zu betonen, bereits bei der Texteingabe berücksichtigt werden. Es muß lediglich eine andere Korrespondenzregel bzw. Gliederungsregel dem NUMMER-Programm eingegeben werden, wie z.B. "£" als Abgrenzungssymbol zweier logischer Sätze. Das Programm NUMMER bereitet mit anderen Worten den Text mit dieser Korrespondenzregel eindeutig nach logischen Sätzen getrennt und numeriert auf. Die weitere Verarbeitung des Textes berücksichtigt nur noch die Interaktionen (III), der Rest der Datei wird nicht abgespeichert.

### 2.3.2 Speicherplatz

Der generell benötigte Speicherbedarf für die Texteingabe (Kap. 2.2.2) läßt sich nur schwer bestimmen, da die interne Organisation eines File- oder Dateisystems anlagen-spezifisch ist. Für das Betriebssystem DEC 1091 ergab sich als Erfahrungswert für ein durchschnittlich langes Interview mit 50 Seiten Ausgabetext ein Speicherbedarf von ca. 160 Blöcken. Ein Block bedeutet hier 128 (36-bit) EDV-Worte. Übertragen auf Texte lassen sich 5 Zeichen (Characters) in

---

18) Kops, M. (1977) S. 54 ff

einem solchen EDV-Wort abspeichern (19).

Für Großrechenanlagen stoßen diese Werte an keine Speichergrenzen (20). Jedoch eine Parallelverarbeitung von über 15 Interviews (also ohne Zwischenspeicherung auf Magnetband) zeigt schon hier die begrenzte Anwendbarkeit für Kleinrechner. Ziel einer programmtechnischen Lösung muß also sein, den Interviewtext so ökonomisch wie möglich abzuspeichern, was bedeuten soll, nicht mehr Platz in Anspruch zu nehmen, den auch ein gewöhnlicher EDITOR benötigen würde.

### 2.3.3 Zugriffszeit

Ein weiteres Problem stellt die Zugriffszeit dar, d.h. die Zeit, die vergeht bis eine gewünschte Interaktion vom System gefunden und ausgegeben (21) wird. Gängige sozialwissenschaftliche Programmpakete zur statistischen Auswertung von z.B. Umfragedaten arbeiten eine Datenmenge immer noch sequentiell von Fall 1 bis Fall n ab, was bedeutet, je grösser eine Datenmenge ist, um so länger dauert der Zugriff auf Fall n.

---

19) Daraus folgt: ein Interview enthält ca. 100.000 Buchstaben (bzw. Zeichen) (160 Blöcke \* 128 Worte \* 5 Zeichen = 102.200 Characters) Bei Verwendung von ASCII-Zeichen (d.h. der Zeichenumfang umfasst Gross- Kleinbuchstaben, Zahlen und Sonderzeichen) muss mit einer Mindestbitzahl von 7 pro Character ausgegangen werden.

20) Eine Speicherplatte des DEC-1091 Betriebsystems fasst 300.000 bzw. 800.000 Blöcke (netto)

21) 'ausgegeben' bedeutet hier entweder auf Drucker oder Bildschirm

Eine solche Dateiorganisation ist für das obige Problem aus 2 Gründen ungeeignet. Erstens darf für ein sinnvolles Arbeiten eine Ausgabe der Interaktionen 5-10 nicht länger dauern, als die von 505-510. Zweitens bauen SPSS-ähnliche Dateiorganisationen rechteckige Filestrukturen auf, da bei großen Datenmengen praktisch nur mit INPUT FORMAT FIXED abgespeichert wird, was bei variabler Satzlänge zur Konsequenz hätte: Ich muß wissen, wieviel EDV-Worte mein längster logischer Satz (Interaktion) benötigt und definiere mit dieser Information das Format, was aber zur Folge hat, auch die kleinste Interaktion (z.B. I: hmm) erhält genauso viel Speicher reserviert, wie eine 3 minütige ununterbrochene Antwort des Befragten. Die Konsequenz wäre zum einen riesiger Speicherbedarf, und zum anderen verlängerter Zugriff bei sequentieller Dateiorganisation.

#### 2.4 Programmtechnische Lösung

Für das Problem der Zugriffszeit bei großen Dateien bietet der RANDOM ACCESS MODE (RAM) (22) eine Lösungsmöglichkeit. Der RAM wurde speziell für COBOL-Compiler konzipiert, d.h. für kommerzielle Anwendungsgebiete, ist aber auch für FORTRAN mittlerweile verfügbar. RAM bedeutet Direktzugriff auf Blöcke mit Speicherworten, die auf Plattendateien organisiert sind.

Beim Speicherplatzproblem wurde versucht eine ebenfalls im kommerziellen Bereich entwickelte Methode zur Verwaltung von Massendaten, auf Textprobleme zu übertragen; die sog. indexsequentielle Datei-Organisation. Die Logik die dahintersteckt ist folgende: Eine RAM-Datei wird durch ein Index-Register verwaltet, wobei durch sog. Record-Schlüssel die physikalische Record-Adresse von logischen Sätzen

---

22) eine ausführliche systemspezifische Beschreibung siehe Digital Equipment Cooperation (1975b), Kap.10-1

identifizierbar und damit direkt zugreifbar wird.

Das vom Autor in FORTRAN entwickelte Programm INDEX baut ein solches Index-Register als 3-dimensionale Matrix auf und speichert die Interviews im RAM ab. Es bleibt freilich anzumerken, daß z.B. IBM ein systemspezifisches indexsequentiellles Dateisystem ISAM (index sequential access method) anbietet, wo Verwaltungs- und Datenteil in einer Datei organisiert sind. Die hier programmierte Lösung ist jedoch kompatibler, da unabhängiger vom jeweiligen Dateisystem.

Die RAM-Datei mit den Transkriptionstexten ist nun nicht mehr von einem normalen EDITOR zu lesen, sondern nur noch vom TMS 'entschlüsselbar'.

#### 2.4.1 Aufgaben eines Textmanagementsystems (TMS)

Als TMS wurde eine Sammlung von Programmen bezeichnet, die zum Ziel haben, maschinenlesbare Texte für Analysezwecke zunächst in Systemdateien überzuführen, die anschließend von einem Interviewverarbeitungsprogramm gelesen werden können, und anschließend mittels verschiedener Programmmodule spezifische Ordnungssysteme für Textsegmente eines - oder auch mehrerer Interviews herstellbar sind. Weiterhin soll das Verarbeitungs- bzw. Verwaltungsprogramm so angelegt sein, daß es leicht durch zusätzliche Programmmodule ergänzbar ist, um weiterführende Analysen durchzuführen.

Die Programme wurden alle in FORTRAN-10 (DEC-1091) geschrieben. Es wurde versucht aus Gründen der Kompatibilität möglichst wenig systemspezifische Programmelemente zu verwenden, was sich leider nicht gänzlich vermeiden lies, da die hier benutzte Fortran-Version noch nicht ANSI FORTRAN 77 entspricht, jedoch mit Character-Variablen operiert werden mußte, die lediglich mit Hilfe standardisierter Assembler-

routinen (nicht vom Autor programmiert sondern in Unterprogramm-bibliothek des Herstellers verfügbar ) umgangen werden konnten.

Das INDEX-Programm versucht mit minimalem Speicherbedarf und Direktzugriff den transkribierten Text, der das Programm NUMMER durchlaufen hat, so aufzubereiten, daß jede gewünschte Interaktion nachher vom TMS bereitgestellt und mit beliebig anderen Interaktionen zusammengefügt werden kann. Die Lösung sollte so programmiert sein, daß eine Bearbeitung mit dem Interviewmaterial interaktiv geschehen kann, also ohne hohen Kernspeicherbedarf des Programms (CPU-intensiv) und benutzerfreundlich in der wahlfreien Eingabe verschiedener Variablen (z.B. der Interviewnummern).

Es sollten dabei die spezifischen Vorteile einer EDV-Anlage auch für die Auswertung qualitativer Interviews und die hierbei gewöhnlich anfallende Textflut genutzt werden. Nämlich neben den Vorteilen der leichteren Korrigierbarkeit bei der Texterfassung, jetzt vor allem die Speichermedien zu nutzen, um einmal eingegebene Texte in beliebiger inhaltlicher Zusammenstellung 'auseinanderzuschneiden' und wieder 'zusammenzusetzen'; Vorgänge die nun ohne Kopierer, Schere und Klebstoff zu erledigen sind.

## 2.5 Zum Umgang mit dem TMS

Im folgenden soll die praktische Arbeit mit den bislang entwickelten und erprobten Auswertungsmodulen beschrieben werden.

### 2.5.1 Auswertungspraktische Ziele des TMS

Sinn und Zweck eines Textmanagementsystems sollte sein, nicht in der nach der Transkriptionsphase entstandenen Textflut von mehreren hundert Seiten zu ersticken, sondern

praktische Arbeitsschritte zu entwickeln, welche die inhaltliche Vielfalt, die in den Texten stecken - also das 'qualitative Substrat' - nicht letztlich doch nur in inhaltsleere Kategorienkästchen einzusortieren.

Eine Verdichtung des Materials durch 'selektive Protokollierung' (23) erscheint problematisch da ein solches Vorgehen z.B. Nachinterpretationen zu einem späteren Zeitpunkt, bzw. durch andere Forscher ausschließt.

Wertvolle Anhaltspunkte für eine programmtechnische Umsetzung bot das von Mühlfeld u.a. (24) dargestellte mehrstufige Auswertungsmuster.

Die Auswertung zielte bei der hier zugrundeliegenden Anwendung zum einen auf ein inhaltliches Füllen, der im Interviewleitfaden eingeführten Themen- und Problemkomplexe der Untersuchungsgruppe, zum anderen auf eine ständige Erweiterung des vom Befragten eingeführten Themen- und Problemerkatalogs, was freilich stets von der Kompetenz und Erzählbereitschaft der Befragten abhängig bleibt. Im Fokus der Betrachtung stand keine Fallrekonstruktion, sondern die Absicht Gemeinsamkeiten in den Aussagen der Informanten herauszuarbeiten, um daraus Typologien im Hinblick auf verschiedene Untersuchungsdimensionen zu entwickeln, welche am Ende in eine sinnvolle Beziehung zu Arbeitshypothesen gesetzt werden können.

---

23) siehe Brand, K.-W. und Honolka, H. (1981), S. 323 f, hier wurde das Problem der Datenreduktion durch 'Vorinterpretation' der Interviewer gelöst, deren Qualität freilich später nicht mehr zu kontrollieren ist.

24) siehe Mühlfeld, C. u.a. (1981), S. 335 ff

Das TMS ist konzipiert als praktische Hilfe für das Zuordnen von Textpassagen zu Problemfeldern und genügt darüberhinaus dem Anspruch, die entsprechenden Textstellen bis zum letzten Arbeitsschritt schnell, vollständig und - wenn gewünscht - auch im Kontext der jeweils spezifischen Fragenkonstellation verfügbar zu halten. Mit anderen Worten, das Material bleibt einer intersubjektiven Prüfung zugänglich. Die Zuordnung soll darüberhinaus leicht korrigierbar und ergänzbar sein.

### 2.5.2 Topics

Der 1. Arbeitsschritt im Umgang mit dem Interviewmaterial bildete wie auch bei Mühlfeld die Identifizierung zentraler, im Interviewleitfaden aufgestellter Problemfelder; hier an Hand der Interaktionsnummer im Interviewtext. Als zweites die von den Befragten selbst, über den Leitfaden hinaus entwickelten Themen ebenfalls zu lokalisieren.

Praktisch bedeutete dies, daß beim erneuten (25) Durchlesen der 'Reinschrift', die einzelnen Interaktionen einem sukzessive erweiterten Kategoriengerüst zugeordnet wurden. Diese erste Codierung der Interaktionen in Topics wurde dem Umstand gerecht, daß eine Äußerung zu einem Thema erst einen Sinn oder auch "höhere Validität" (26) erreicht,

---

25) Der Autor war mit dem Interviewtext dann zum 5. mal in seiner kompletten Länge konfrontiert, was zu einer spürbaren Sensibilisierung dem Text gegenüber führte

- a) Interviewdurchführung
- b) Rohtranskription
- c) Texteingabe
- d) Korrekturhören

26) vgl. Kohli, M. (1978), S. 20

".. es ist möglich die Ergebnisse einer hermeneutischen Interpretation einer quantitativen Verarbeitung zuzuführen, indem sie - nach Abbruch des Zirkels - auf Variablen verteilt und deren Ausprägungen zugeordnet werden. Die Interpretation wird dann als eine besondere Art von Kodierung aufgefasst, wobei durch den Bezug aller Daten auf den Gesamtkontext eine höhere Validität erreicht werden soll."

wenn die Daten auf den Gesamtkontext bezogen sind. Mit anderen Worten, ein und dieselbe Interaktion konnte mehreren Kategorien zugeordnet werden, wie umgekehrt eine Kategorie erst Sinn geben konnte bei Berücksichtigung von Aussagen in verschiedenen Teilen des Interviews.

Für die Erstellung des Kategoriengerüsts sind 2 Lösungen denkbar.

- a) ein Kategoriengerüst, das für alle (oder zumindestens für eine bestimmte Subgruppe) gültig ist.
- b) ein Kategoriengerüst, das vollkommen fallspezifisch ist.

Beide Möglichkeiten sind EDV-technisch realisierbar, die praktischen Erfahrungen geben jedoch Fall a den Vorzug. Hierfür wurde das Modul KAT des TMS entwickelt, welches mit Hilfe einer editierten Kategorienliste (die ständig erweitert oder modifiziert werden kann) interaktiv die interviewspezifischen Interaktionsnummern zu jedem einzelnen Topic abfragt. Das Programm erstellt dann mit dieser Information eine Datei, welche für weitere Programmschritte notwendig ist.

#### ABBILDUNG TOPIC-Format

NR	TOPIC-Text	Interaktionsnummern	Spaltenzahl
1/	/5/	/61/	
-----////////-----			
..			
..			
11	Probleme als Frau in Beruf	40,120,133-135	
12	selbst verwendete Programmiermethode	300,330-337,428	
..			
20	EDV-Kenntnisse von Vorgesetzten	0	
..			

Erläuterung

1. In Spalte 1-2 befindet sich die laufende Topic-numerierung.
2. Von Spalte 5-60 der Text des jeweiligen Topics (in Fall b interviewspezifisch).
3. Ab Spalte 61 die zugehörigen Interaktionsnummern durch Kommata getrennt, bzw. bei Intervallen durch Gedankenstrich (in Fall a und b interviewspezifisch).

Das Modul ONE des TMS ermöglicht anschließend an Hand der interviewspezifischen (27) Topic-Liste für jedes Interview einen themenzentrierten Ausdruck zu erstellen, der eine andere Ordnung besitzt als der eigentliche Verlauf des Interviews. Hierfür wurden wiederum die Vorteile des Textverarbeitungsprogramms RUNOFF zur übersichtlichen Gestaltung des Ausgabetextes (Interviewcode und jeweilige Kategorie als Kopfzeile) mit dem Programmmodul verknüpft.

Eine solche aufbereitete (oder problemorientierte) Interviewversion kann nun für die weitere Interpretationsphase genutzt werden. Für das Programm selbst spielt es dabei keine Rolle, ob eine Interaktion, da sie z.B. über mehrere Zeilen geht, inhaltlich mehreren Topics zugeordnet wurde; ebenso gibt es keine Beschränkungen in Bezug auf die Anzahl der Topics.

Für Topics, die nur in wenigen Interviews auftauchen, also eher fallspezifisch sind, wurde die Möglichkeit eingebaut, daß das Programm ein Fehlen dieses Topics bei anderen Interviews anzeigt (28).

- 
- 27) soll hier bedeuten, dass das gleiche oder ähnliche Topic in jedem Interview zu einem beliebigen Zeitpunkt im Interviewablauf (symbolisiert durch die Interaktionsnr.) auftauchen kann.
  - 28) d.h. wenn in Spalte 61 eine 0 steht (vgl. Abb. Topic-Format, Nr. 20) erscheint im Ausdruck:  
In Interview XXX fehlt Topic  
"EDV-Kenntnisse von Vorgesetzten"

Ein Vorteil in der Anwendung der EDV käme auch dann zum Vorschein, wollte man die themenzentrierten - bzw. Kategorien zugewiesenen - Interviewpassagen reinterpretieren, bzw. wollte man in einer 2. Stufe der Interpretation bei Wiederholungen/Ähnlichkeiten jetzt nur die jeweils prägnanteste Textstelle zugrunde legen. In diesem Fall müßte man nicht wieder von vorn mit der Arbeit beginnen; der einzige Arbeitsaufwand der entstünde, wäre die Korrektur einer Topic-Liste mit dem Editor.

### 2.5.3 Summaries

Nachdem für jedes Interview der Inhalt einzelnen Themen oder Kategorien zugeordnet war, sollte für die weitere Auswertung, bzw. Entwicklung von Typen eine programmtechnische Lösung entwickelt werden, problemgeordnete Zusammenfassungen aller Interviews zu erstellen. Informationsgrundlage eines solchen Screenings über mehrere Interviews bilden wiederum die Topics. Bündelt man nun mehrere Topics zu einem Themenschwerpunkt, so stellt das Programmmodul SUM integriert mit dem Textverarbeitungssystem RUNOFF folgenden Aufbau (29) eines Ausdrucks auf.

1. Kopfzeile: Themenschwerpunkt
2. Kopfzeile: Einzeltopic  
darunter Interviewmaterial, das inhaltlich zu obigen Kategorien zugewiesen wurde. Unterbrochen vom jeweiligen zentrierten Interviewcode.

Das Programmmodul läuft wiederum interaktiv, indem zunächst nach der Zahl der zu verarbeitenden Interviews 'gefragt' wird, anschließend nach dem Themenschwerpunkt und als letztes nach den Topicnummern, die unter die 'Summary' gefaßt sind. Anschließend durchläuft das Programm 'relativ rasch' (30) sämtliche Interviews (oder nur eine spezifische

---

29) jeweils durch die Anwendung von RUNOFF Steuerzeichen

30) relativ soll bedeuten, dass das Programm im DEMAND-Betrieb mit wenig Hauptspeicher-Bedarf zu nutzen ist

Auswahl). Der erstellte Ausdruck eignet sich nun als Arbeitsgrundlage für weitere Interpretationen und Typisierungen des Materials. Die leichte Modifizierbarkeit von Ordnungssystemen und die Möglichkeit der Subgruppenbildung erweist sich wahrscheinlich erst bei größeren Forschungsteams als große Hilfe, wenn Interpretationen sukzessive und arbeitsteilig erarbeitet werden.

#### 2.5.4 weitere Programmfunktionen

Das TMS wurde bei der vorgenommenen Auswertung des Materials noch um das Dokumentationsmodul DOC erweitert. Da für die Beschreibung der Ergebnisse der Befragung auf den Originaltext Bezug genommen werden sollte, wurde eine Form der Darstellung gewählt, die zum einen bei der reinen Deskription der Ergebnisse auf die Beschreibungskompetenz der Informanten zurückgreift, zum anderen sich eine Interpretation bestimmter Sachverhalte an dem extrahierten Material kontrollierbar ausweisen ließe.

Hierzu wurden am linken Rand der Dokumentation der Ergebnisse jeweils die 'Fundstellen' im Textmaterial festgehalten. Mit diesem 'Code' wurde ein Anhang erstellt, der die jeweiligen Quellen in der Reihenfolge ihrer Erwähnung im Auswertungsteil enthält. Erstellt man die Dokumentation der Ergebnisse ebenfalls EDV-gestützt, so läßt sich die praktische Erstellung dieses Quellenanhangs gar noch automatisieren. Auch an diesem Punkt zeigten sich wieder die Vorteile einer Verknüpfung mit der Standardsoftware der Textverarbeitung, da sich durch wenig Parameteränderungen eine andere Präsentation des Transkriptionsmaterials erzeugen ließ. Mit Hilfe dieser Belegführung der Ergebnisse läßt sich sicherlich die intersubjektive Überprüfbarkeit qualitativer Studien erhöhen.

Mit dem DOC-Programm lassen sich interaktiv nach Eingabe der Ausgabeeinheit (hier ist auch der Bildschirm als Ausgabemedium vorgesehen) das Interview und dann hierin jede beliebige Interaktion (bzw. -intervall) nochmals vor Augen führen oder ausdrucken, was das Material und vor allem auch die vorgenommenen Interpretationen einer intersubjektiven Prüfung zugänglich macht, da jeder Benutzer des Systems sich seine Interviewpassagen individuell zusammenstellen kann, ohne dabei auf den jeweiligen 'Gesamtanlageband' mit der Reinschrift aller Interviews zurückgreifen zu müssen.

Als mögliche Ergänzung zum bisher entwickelten TMS wäre denkbar, die Topic-Files, welche das interviewspezifische Auftreten einzelner Forschungsfragen widerspiegeln, mit 'quantitativen' Modulen zu koppeln und deskriptiv-statistische Verteilungsmaße zu bestimmten Kategorien aufzustellen. Sinnvoll wird dieses Vorgehen freilich erst bei einer genügend hohen Anzahl von erfaßten Interviews.

### 3.0 ZUSAMMENFASSUNG DER VORTEILE VON EDV

- flexiblere und schnellere Anpassung von Änderungen (Textkorrektur, unterschiedliches Ausgabeformat)
- schnelleres Erstellen von Ordnungssystemen
- größere Transparenz des qualitativen Materials (Transkripte)
- Möglichkeit der gleichzeitigen Nutzung des Erhebungsmaterials durch mehrere Forscher
- Unterstützung der Möglichkeit sukzessive Reinterpretationen vorzunehmen.
- Wegfall von Schreibearbeit (für Dokumentation)
- und vielleicht auch geringere Kosten (1000 Blatt weißes Endlospapier wird in Frankfurt mit DM 36,-- berechnet).

LITERATURVERZEICHNIS

- Brand, K.-W. und Honolka, H. (1981)  
Lebenswelt und Wahlentscheidung. Eine explorative Studie zur Bundestagswahl 1980 mit Hilfe von Bekannteninterviews, in: PVS, Jg. 22, Heft 3, S. 305-317
- Digital Equipment Cooperation (1975a)  
RUNOFF, in: Decsystem 10 Utilities Manual, Maynard (Mass), S. 187-233
- Digital Equipment Cooperation (1975b)  
FORTRAN-10 Programmer's Reference Manual, Maynard (Mass)
- Dworatschek, S. (1977)  
Grundlagen der Datenverarbeitung, 6. Auflage, Berlin - New York
- Hall, Ch. (1981)  
SED: A Screen Editor for TOPS 10. Programmer's Reference Manual (o.O.) (Programmdokumentation über Digital Equipment Co.)
- Kohli, M. (1978)  
'Offenes' und 'geschlossenes' Interview: Neue Argumente zu einer alten Kontroverse, in: Soziale Welt, Jg. 29, Heft 1, S. 1-25
- Kops, M. (1977)  
Auswahlverfahren in der Inhaltsanalyse, Meisenheim
- Kriz, J. (1975)  
Datenverarbeitung für Sozialwissenschaftler, Reinbeck
- Muehlfeld, C. ; Windolf, P. ; Lampert, H. und Krueger, H. (1981)  
Auswertungsprobleme offener Interviews, in: Soziale Welt, Jg. 32, Heft 3, S. 325-352
- Schupp, J. (1982)  
Zur Arbeits- und Berufssituation von Mathematisch-Technischen Assistenten(innen) - Eine empirische Fallstudie - (mit EDV-gestützter Auswertung qualitativer Daten), Diplomarbeit am Fachbereich 3 Gesellschaftswissenschaften der Joh. Wolfgang Goethe Universität Frankfurt
- Stone, J.Ph. u.a. (1966)  
The General Inquirer: A Computer Approach to Content Analysis, Cambridge (Mass) - London
- Wossidlo, P.R. (Hrsg.) (1980)  
Textverarbeitung und Informatik: Fachtagung der GI, Bayreuth, 28.-30. Mai 1980, Berlin - Heidelberg - New York

INTERVIEW 77/9-A-B

SEITE 13

- 00415 B: DASS ES DA AUCH NOCH ANDERE MOEGlichkeiten GIBT 'NE, ABER DIE DAS IST EBEN DAS GUTE DRAN, DASS DIE LEUTE AUCH BEREIT SIND, AUCH MAL WAS AUFZUNEHMEN, NE
- 00416 I: HMM
- 00417 B: ODER WAS NEUES ANZUNEHMEN, DER EINE HAT SOGAR MAL MEINE UNTERLAGEN AUSGELIEHEN UND SO UND SO DAS FAND' ICH ECHT GANZ TOLL, ICH DACHT' NAEMLICH SCHON, DASS SIE JETZT KOMMEN, ACH ICH PROGRAMMIER SCHON 10 JAHRE SO UND WARUM SOLL DAS ANDERE BESSER SEIN UND DU MUSST DAS SO MACHEN WIE WIR DAS HABEN WOLLEN, ANSONSTEN IST DAS NICHT GUT UND SO, DIE HABEN GESAGT, AH JA DAS STIMMT, DAS FUNKTIONIERT AUCH SO, IST EIGENTLICH GAR NICHT SO SCHLECHT, JE NACHDEM
- 00418 I: HMM, ABER NORMAL WERDEN DIE PROGRAMME EH, WEIL DU GESAGT HAST DU MACHST LIEBER WAS, WO DU SELBST MACHEN MUSST UND NET IRGENDWAS VERBESSERN
- 00419 B: JA, JA
- 00420 I: AUCH DESHALB WIRD DIR HALT, MIT 'NER ANDEREN MENTALITAET GESCHRIEBEN WORDEN SIND
- 00421 B: GENAU, JA, DAS IST
- 00422 I: ALS IHR DAS GEMACHT HABT
- 00423 B: BRAUCH ICH VIEL LAENGER BIS ICH IM PROGRAMM DRIN BIN, ALS WIE ICH MEIN EIGENES GESCHRIEBEN HAB', IST JA GANZ KLAR, NE, WENN DA STEHT GO HIERHIN, DANN SPRINGT ER WIEDER GANZ OBEN HIN UND DANN SETZT ER DA 'EN SCHALTER UND DANN SPRINGT ER WIEDER WEG UND ICH WEISS JA AUCH NET FUER WAS DIE SCHALTER DA SIND, NE
- 00424 I: JA
- 00425 B: DAS IST MEISTENS SO SCHLECHT DOKUMENTIERT, DASS DA NIRGENDS STEHT FUER WAS WIRD DER GESETZT ODER EH', WENN ICH DANN WAS AENDERE, HAU' ICH MIR VIELLEICHT DADURCH WAS KAPUTT BEI DEM, WEIL ICH DAS PROGRAMM AUCH NICHT GENAU DURCHSCHAU, NE
- 00426 I: HMM
- 00427 B: DA MUSS ICH MEIN, ICH HAB' MIR HEUT' MAL DIE MUEHE GEMACHT, HAB' MIR MAL DIE VORGABE NOCH DURCHGELESEN, ABER DAS HAT MIR AUCH NET SO SEHR VIEL GEBRACHT, NE, WEIL DAS EINPACH DAS PROGRAMM AN SICH ZU GROSS WAR UND DIE VORGABEN DA DIE SIND AUCH NET SO SEHR AUSFUEHRLICH, ALSO ES KOMMT AUCH DRAUF AN, WIE DIE VORGABEN GESCHRIEBEN SIND,
- 00428 I: HMM
- 00429 B: MANCHE SIND SEHR GUTE, ANDERE WIEDERUM DIE BIND SO DUERFTIG, DASS MAN GAR NET WEISS, WAS MAN EIGENTLICH MACHEN SOLL, WENN MAN NET GRAD DAS HINTERGRUNDWISSEN HAT, WAS MAN, WAS ABER MEISTENS FEHLT, NE

- 00415 B: dass es da auch noch andere Moeglichkeiten gibt 'ne, aber die das ist eben das gute dran, dass die Leute auch bereit sind, auch mal was aufzunehmen, ne
- 00416 I: hmm
- 00417 B: oder was neues anzunehmen, der eine hat sogar mal meine Unterlagen ausgeliehen und so und so das fand' ich echt ganz toll, ich dacht' naemlich schon, dass sie jetzt kommen, ach ich programmier schon 10 Jahre so und warum soll das andere besser sein und du musst das so machen wie wir das haben wollen, ansonsten ist das nicht gut und so, die haben gesagt, ah ja das stimmt, das funktioniert auch so, ist eigentlich gar nicht so schlecht, je nachdem
- 00418 I: hmm, aber normal werden die Programme eh, weil du gesagt hast du machst lieber was, wo du selbst machen musst und net irgendwas verbessern
- 00419 B: ja, ja
- 00420 I: auch deshalb wird dir halt, mit 'ner anderen Mentalitaet geschrieben worden sind
- 00421 B: genau, ja, das ist
- 00422 I: als ihr das gemacht habt
- 00423 B: brauch ich viel laenger bis ich im Programm drin bin, als wie ich mein eigenes geschrieben hab', ist ja ganz klar, ne, wenn da steht 30 hierhin, dann springt er wieder ganz oben hin und dann setzt er da 'en Schalter und dann springt er wieder weg und ich weiss ja auch net fuer was die Schalter da sind, ne
- 00424 I: ja
- 00425 B: das ist meistens so schlecht dokumentiert, dass da nirgends steht fuer was wird der gesetzt oder eh', wenn ich dann was aendere, hau' ich mir vielleicht dadurch was kaputt bei dem, weil ich das Programm auch nicht genau durchschau, ne
- 00426 I: hmm
- 00427 B: da muss ich mein, ich hab' mir heut' mal die Muehe gemacht, hab' mir mal die Vorgabe noch durchgelesen, aber das hat mir auch net so sehr viel gebracht, ne, weil das einfach das Programm an sich zu gross war und die Vorgaben da die sind auch net so sehr ausfuehrlich, also es kommt auch drauf an, wie die Vorgaben geschrieben sind,
- 00428 I: hmm
- 00429 B: manche sind sehr gute, andere wiederum die sind so duerftig, dass man gar net weiss, was man eigentlich machen soll, wenn man net grad das Hintergrundwissen hat, was man, was aber meistens fehlt, ne

### 3. Spezielle Anwendungen:

#### Halboffene Verkodung als Mittel der Datenreduktion

Der hier wiedergegebene Beitrag aus den ZUMA-Nachrichten Nr.12 befaßt sich mit der Möglichkeit, bei großen Mengen sprachlichen Materials nur einen geringen Teil maschinenlesbar aufzubereiten. Neben der klassischen Verkodung sprachlichen Materials, i.e. der Zuweisung von Ziffern zu Textmerkmalen, wird die Möglichkeit der halboffenen Verkodung besprochen. Das Prinzip dieser Verkodung besteht in der Verbindung von Schlagworten bzw. Schlüsselworten und den ihnen zuzurechnenden Volltexten.

Dies Verfahren ist mit J. Schupps Verschlagwortung (Teil 3) vergleichbar, nur werden dort zuerst alle Texte maschinenlesbar aufbereitet und dann verschlagwortet; hier hingegen werden nur Textstellen, denen ein Schlagwort zugewiesen werden kann, maschinenlesbar aufbereitet.

## PROJEKTBERICHT: SYSTEMATISCHE AUFARBEITUNG DES ARCHIVS DES SIGMUND-FREUD- INSTITUTS IN FRANKFURT

### 1. Einleitende Bemerkungen

Im Januar 1981 trat das Sigmund-Freud-Institut (SFI) an ZUMA mit der Bitte um Unterstützung bei der Auswertung seines Archivs heran, in dem die sogenannten Ambulanz-Unterlagen gespeichert sind. Dabei handelt es sich um die schriftlichen Unterlagen, die im Verlauf und nach Abschluß der diagnostischen Untersuchung eines Patienten, der das Sigmund-Freud-Institut aufsucht, angefertigt werden.

Das Sigmund-Freud-Institut ist ein Forschungs- und Ausbildungsinstitut für Psychoanalyse.\* Es hat im wesentlichen folgende vier Aufgabenbereiche:

1. Erforschung der Lebens- und Krankheitsbedingungen der Menschen mit Hilfe der psychoanalytischen Methode und in Zusammenarbeit mit angrenzenden humanwissenschaftlichen Disziplinen.
2. Weiterbildung von Ärzten und Psychologen zu Psychoanalytikern (nach den Richtlinien der Deutschen Psychoanalytischen Vereinigung) sowie Untersuchung didaktischer und methodischer Fragen zur Weiterbildung.
3. Psychoanalytische Behandlung von Patienten. Bei der Auswahl der Patienten sind Forschungs- und Ausbildungsgesichtspunkte zu berücksichtigen.
4. Vermittlung psychoanalytischer Erkenntnisse an Angehörige einschlägiger Berufe und Entwicklung von Methoden und Modellen, die die Anwendung psychoanalytischer Erkenntnisse in diesen Fachgebieten fördern.

Zur Erfüllung seiner Aufgaben unterhält das Sigmund-Freud-Institut eine Ambulanz, die von Patienten auf eigenen Entschluß oder als Folge einer Überweisung aufgesucht werden kann. Die Aufgabe der Ambulanz ist in erster Linie eine diagnostische. Aufgrund der großen Zahl von Patienten, die das Institut aufsuchen, und aufgrund der vielfältigen Aufgaben der im Sigmund-

---

\*Das Institut wird vornehmlich vom Land Hessen finanziert.

Freud-Institut tätigen Psychoanalytiker kann nur einer Minderheit der Patienten eine Behandlung im Institut angeboten werden. Unterlagen über den Verlauf solcher Behandlungen werden - wenn man von Kurzbehandlungen oder Beratungen absieht - in dem zur Frage stehenden Archiv nicht gesammelt.

Das Archiv enthält eine Fülle klinischen Materials, das bis jetzt nicht systematisch ausgewertet wurde. Es gibt auch keine andere vergleichbare systematische Untersuchung dieser Größenordnung bei einer Patientenklientel wie der, die das Sigmund-Freud-Institut aufsucht. Daraus ergibt sich die Bedeutung der Untersuchung für psychoanalytische, aber auch sozialpsychologische Fragestellungen.

Im folgenden werden das Archiv beschrieben, die Ziele des Projekts vorgestellt, der allgemeine Lösungsansatz skizziert und an einigen Beispielen konkretisiert.

## 2. Das Ambulanzarchiv und die Forschungsziele der Projektgruppe

Seit seiner Gründung im Jahre 1960 führt das SFI diagnostische Untersuchungen durch und bietet in beschränktem Umfang auch Behandlungen an. Patienten, die das Institut aufsuchen, füllen zunächst ein kurzes Formular aus. Wie alle Formulare, die im Archiv angelegt sind, unterlag auch dieses im Laufe der Zeit erheblichen formalen und inhaltlichen Veränderungen. Der Patient wird dann einem Analytiker zugewiesen, mit dem er in der Regel ein kurzes Vorgespräch führt. Bei einem Teil der Patienten bleibt es bei diesem kurzen Vorgespräch, entweder, weil sich herausstellt, daß der Patient nicht mehr als eine kurze Beratung braucht, oder weil sich herausstellt, daß er mit seinen Beschwerden besser eine andere Institution aufsuchen sollte. Der Mehrzahl der Patienten wird zur weiteren Abklärung ein diagnostisches Interview angeboten. Dieses ist der zentrale Teil der Untersuchung. In einem Teil der Fälle stellt sich beim Vorgespräch heraus, daß zwar keine diagnostische Untersuchung, wohl aber eine eingehendere Beratung indiziert ist. Dann findet anstelle des diagnostischen Interviews ein Beratungsinterview statt. Vom diagnostischen Interview, das gelegentlich wiederholt wird (beim selben Analytiker oder bei einem anderen) wird ebenso wie vom Beratungsinterview nach der Sitzung ein Gedächtnisprotokoll angefertigt. Das Protokoll, das vom diagnostischen Interview angefertigt wird, enthält außer der

# ZUMA

---

Beschreibung des Interviewverlaufs auch die Diagnose und den Therapievor-schlag des Analytikers. Auf letztere Angaben, nämlich die zur Diagnose und zur Therapie, kann im Beratungsinterview verzichtet werden. Während mit ei-nem oder möglicherweise auch mehreren Beratungsinterviews der Untersu-chungsgang abgeschlossen ist, folgt auf ein diagnostisches Interview in vielen Fällen noch ein psychologischer Test. Auch wenn ein solcher Test nicht durchgeführt wird, ist es üblich, daß das Gedächtnisprotokoll der so- genannten Ambulanzkonferenz zur Diskussion vorgelegt wird. Wenn eine test- psychologische Untersuchung durchgeführt wird, werden selbstverständlich auch die Testunterlagen dieser Konferenz zugänglich gemacht. Die Ambulanz- konferenz kommt dann zu einer abschließenden Beurteilung des Patienten, und abschließende Diagnose und Behandlungsvorschlag werden vom betreuenden Ana- lytiker in einer kurzen Notiz festgehalten.

In einem Abschlußgespräch, von dem es zumeist nur eine kurze Notiz gibt, werden dem Patienten der Therapievor-schlag erläutert und mögliche Maßnahmen besprochen. Außer den Notizen, Formularen und Protokollen finden sich ge- gentlich Briefe und andere Unterlagen in den Akten.

Das Archiv umfaßt ca. 3600 Fälle, von denen einige nicht zur Zielpopulation gehören. Hierunter fallen insbesondere Akten von Patienten bzw. Probanden, die für wissenschaftliche Studien ins SFI gebeten wurden. Für die Zeit zw- ischen 1960 und 1980 ist die Grundgesamtheit deshalb nur ungefähr angebbar; erst im Verlauf der Datenerhebung kann festgestellt werden, ob ein Fall zur Zielpopulation zu zählen ist.

Geht man pro Fall von ca. zehn Blättern Dokumentation aus, so ergibt sich für das Archiv bis 1980 ein Umfang von ca. 30.000 Blättern, die, wie schon zuvor erwähnt, in Form und Inhalt erheblich variieren.

Das Forschungsziel besteht darin, das heterogene Material systematisch auf- zubereiten, um verschiedene Zielvorstellungen realisieren zu können. Die Deskription der Patientenpopulation über den Zeitraum von 1960 bis 1980 ist das erste und vordringliche Ziel der Projektgruppe. Diese Deskription soll sowohl die demographische Beschreibung als auch die Darstellung von Sympto- matik und Diagnose umfassen. Ein weiteres Ziel ist die Überprüfung bestimm- ter Hypothesen, die in der Literatur über die psychoanalytische Diagnostik

geäußert werden. Auch soll eine Datenbasis geschaffen werden, die es Forschern mit anderen Problemstellungen als denen der Projektgruppe erlaubt, Analysen durchzuführen. In all den Bereichen sollen Längsschnitts-, Querschnitts- und Fallstudien möglich sein.

### 3. Der allgemeine Lösungsansatz

#### 3.1 Kumuliertes Sampling

Vorgesehen war, alle Fälle, die zur Zielpopulation zu zählen sind, vollständig zu erfassen. Nach den ersten Erfahrungen mit dem unten erläuterten Erfassungsschema war offensichtlich, daß innerhalb des Zeitrahmens von einem Jahr, der für die Erfassung angesetzt war, nur ein Teil der Fälle bearbeitet werden konnte. Deshalb teilten wir die Grundgesamtheit in 70 Zufallsstichproben (ohne Zurücklegen) à 50 Fälle. Jede Stichprobe streut über den gesamten Untersuchungszeitraum. Die einzelnen Stichproben werden nach und nach abgearbeitet und zu einer Gesamtstichprobe kumuliert. Dies Verfahren erlaubt, nach mehreren hundert Fällen an einer beliebigen Stelle abzubrechen und dennoch eine repräsentative Stichprobe zur Verfügung zu haben. Damit werden die Verocoder von einem unziemlichen Zeitdruck befreit und somit die Vercodungsqualität erhöht.

#### 3.2 Aufteilung der Information in a priori und ex posteriori Kategorisierung

Im Prinzip ist eine standardisierte, auf konventioneller Vercodung basierende Datenerfassung auch für ein Projekt vom Umfang der "psychoanalytischen Ambulanz" das Verfahren der Wahl. Allerdings setzt die konventionelle Vercodung der Verwendung der Daten enge Grenzen. In jedem Fall verlangt eine derartige Vercodung eine a priori Kategorisierung, die rigide für den gesamten Erfassungsprozeß durchzuhalten ist. Dies setzt u. a. voraus, daß schon bei der Entwicklung des Kategorienschemas ein genaues Analysedesign vorliegt, dessen Ansprüche dann in der Kategorisierung eingelöst werden müssen. Allein das Ziel, die Datenbasis auch anderen Forschern zugänglich zu machen, verbietet eine alleinige a priori Kategorisierung. Auch gibt es eine Reihe von Problemen, für die eigentlich erst ex post Kategorien zu finden sind. Dazu gehören z. B. die Diagnosen und die Symptome, weil derer

Zuordnung von einer ganzen Reihe anderer Informationen abhängen kann.

Das entgegengesetzte Verfahren zur konventionellen Datenerfassung ist die vollständige Übertragung der Akten auf Datenträger zusammen mit einigen ordnenden Informationen. Danach hat dann jeder Forscher die ihm genehme Kategorisierung ex posteriori durchzuführen. Wesentliche Nachteile dieses Verfahrens sind der Zeitverbrauch für die Abschrift und der entstehende Informationsballast.

Schätzt man, daß die 30.000 Aktenblätter ca. 20.000 Seiten normalem Text entsprechen, so ergibt sich, bei 4 Seiten pro Stunde Erfassungszeit, daß für die Erhebung aller Fälle mehr als 2 Jahre benötigt würden, ohne daß mit dem so erfaßten Material sofort komplexe Analysen durchführbar wären.

Zugleich würde damit eine Datenmenge erzeugt, deren Verwaltung in keinem Verhältnis zu dem Ballast unbrauchbarer Informationen stünde.

Deshalb entschlossen wir uns in einem zweistufigen Verfahren, die Informationen in "sicher a priori kategorisierbare" und "wahrscheinlich nur ex posteriori kategorisierbare" zu unterteilen. Im ersten Schritt wurde festgelegt, welche Informationen überhaupt erfaßt werden sollten. Nicht erfaßt werden Informationen, die offensichtlich extrem selten oder nur sehr undeutlich in den Akten zu finden sind. Dazu gehören u. a. Beschreibungen des äußeren Erscheinungsbildes des Patienten. Erst nach dieser Entscheidung wurde die Erhebungsform für die einzelnen Merkmale festgelegt. Im Anhang I ist das Erhebungsinstrumentarium aufgelistet, aus dem sich die einzelnen erhobenen Merkmale ableiten lassen.

Für a priori kategorisierbare Merkmale galt die Faustregel, daß es sich um "harte" Tatsachen handeln sollte. Dazu gehörten Informationen wie Interviewdatum, Arten der Vorbehandlung, im SFI durchgeführte Untersuchungen, Wohnsituation, formale Partnerschaftsbeziehungen, Zahl der Kinder usw. (vgl. Anhang I).

Nicht a priori kategorisierbar waren zum einen alle Informationen, die mit Zeitdauerangaben verbunden waren, und zum andern, wie schon erwähnt, Informationen zur Symptomatik und Diagnose. Mit Zeitdauer verbundene Informatio-

nen als ex posteriori kategorisierbare zu definieren, hatte vor allem technische Gründe. Denn bei Daten dieser Art ergibt sich das Problem vieler fehlender Angaben. So ist von vornherein abzusehen, daß es Patienten gibt, die nur wenige Kindheitssymptome angeben werden, während andere eine ganze Fülle solcher Symptome nennen werden. Bei einer konventionellen Erfassung müßten aber Platz (Variablen) für die maximale Zahl der Nennungen geschaffen werden. Nimmt man an, daß maximal zehn Kindheitssymptome angegeben würden, dann müßte für die zweiseitige Vercodung des Merkmals und die jeweilig zweiseitige Vercodung des Anfangs- und Endalters, zusammen 60 Spalten, Platz vorgehalten werden. In der Regel werden davon aber nur 18 Spalten besetzt sein. Kindheitssymptome sind aber nicht das einzige derartige Merkmal. Andere sind "aufgewachsen bei", "Leiden der Eltern", "Wohnungswechsel", "Operationen" usw. Offensichtlich würde hier, bildlich gesprochen, ein Schweizer Käse produziert, der mehr Löcher als Käse enthielte. Und wenn dann doch ein Patient elf Angaben zur Kindheitssymptomatik hätte, wäre das ganze Design durcheinandergebracht worden.

Die Informationen zur Symptomatik und Diagnose sind unserer Ansicht nach nur ex posteriori zu klassifizieren, weil die Daten auch anderen Mitarbeitern des SFI später zur Verfügung stehen sollen. Die Klassifikation solcher Daten hängt nicht nur vom jeweiligen theoretischen Standpunkt ab, sondern variiert auch mit Kontextveränderungen. Deren a priori Kategorisierung hätte ein sehr feingegliedertes Kategorienschema vorausgesetzt, das auch hier in eine "Schweizer-Käse-Datei" ausgeartet wäre.

#### 4. Die Erhebungsinstrumente

Zur Erfassung der a priori und ex posteriori kategorisierbaren Informationen wurden die folgenden beiden Erhebungsinstrumente entwickelt:

##### a) Quasi-Questionnaire

Der Quasi-Questionnaire dient zur Erhebung der a priori kategorisierbaren Informationen. Er ist, wie der Name schon sagt, ähnlich aufgebaut wie der Fragebogen einer standardisierten Umfrage (vgl. Anhang I).

Das technische Problem bei der Entwicklung des Erhebungsbogens bestand in der Anordnung der einzelnen Merkmale zu sinnvollen Blöcken und der

Komprimierung auf engstem Raum, ohne die Übersichtlichkeit einzuschränken.

Da die Informationen, die mit dem Quasi-Questionnaire erhoben werden, in den Akten an den unterschiedlichsten Stellen und in unterschiedlicher Reihenfolge auftreten, ist es für den Vercoder wichtig, daß die generellen Erhebungsthemen (Demographie, Vorbehandlung, Familie etc.) im Erhebungsbogen leicht merkbar angeordnet sind. Dadurch muß er nicht lange suchen, wo er ankreuzen soll, und kann auch leichter prüfen, ob er bestimmte Informationen vergessen hat zu vercoden. Demselben Zweck dient die Komprimierung von 206 Variablen auf drei DIN-A4-Seiten; damit kann der Vercoder den gesamten Erhebungsbogen vor sich ausbreiten und muß nicht während der Vercodung hin- und herblättern.

## b) Quasi-Text

Die ex posteriori kategorisierbaren Informationen werden in verkürzter Form abgeschrieben. Jeder Information wird ein Schlagwort zugeordnet. Die Liste der Schlagwörter ist in Anhang II wiedergegeben. Die Vercoder sind angewiesen, genau so viel Text abzuschreiben, daß der Sinn der Information zusammen mit dem Schlagwort eindeutig ist. Also "OPER(ation) Mandeln 15-15" soll bedeuten, daß der Patient sich im Alter von 15 Jahren einer Mandeloperation unterzog, oder "LEI(den) V(ater) Suizidversuch 18-18", daß der Vater des Patienten, als letzterer 18 Jahre alt war, einen Selbstmordversuch unternommen hatte. Da die einzelnen Textstücke, die gelegentlich auch mehrere Sätze umfassen können, durch die Schlagwörter eindeutig gekennzeichnet sind, kann der Vercoder die Informationen in der Reihenfolge, wie sie in der jeweiligen Akte stehen, transkribieren. In Abbildung 1 ist eine solche Transkription wiedergegeben. Besonders zu beachten ist, daß die Information zu dem Schlagwort PSYC über mehrere Stellen verteilt ist. Bei der Datenaufbereitung werden diese Stellen durch Sortieren nach Schlagwörtern zusammengeführt. Der Schrägstrich bedeutet das Ende einer Sinneinheit.

## 5. Datenaufbereitung

Datenaufbereitung meint die Überführung der Informationen von den Erhebungsbögen in den Computer. Hier werden besonders die Übertragung der In-

# ZUMA

Abb. 1: Transkriptionsbeispiel Quasi-Text

CODEBLATT HALBOFFENER CODE  
für PSYCHOANALYTISCHE AMBULANZ

Vercoder: \_\_\_\_\_ Datum: \_\_\_\_\_ Seite: \_\_\_\_\_

Ambulanznr.	Mnemo	TEXT
9999	WORT	Nannheim /
	BPAT	Bankkauffmann /
	PSYC	Suizidversuch /
	PSOM	Tragereuchtsproblematik /
	AUFW	0-16 beide Eltern /
		17-18 mit Stiefvater /
	EBEZ	0-16 Eltern zusammen /
	LEIV	18-18 Suizidversuch /
	BTUT	Kaufmannsgehilfin
	BVAT	Kaufmann
	PSYC	NØ 3-7 / V6 5-10 /
	DSYI	Tragereuchts
	DSII	oral-anale Problematik /
	OPER	25-25 Mandeln /
	DSI	relativ benigne Tragereuchts- struktur / Trotz, Narziss- mus Reaktionsbildung auf Aggressivität.

Wichtig: für jede neue Ambulanznummer ein neues Codeblatt anlegen!  
kein TEXT in die beiden ersten Felder eintragen!  
Seitenzahlen nicht vergessen! U N D bitte Bemerkungen  
ins allgemeine Codebuch eintragen !

=====

formation auf einen Datenträger (Datenerfassung) und die Manipulation der erfaßten Daten (Datenmanagement) beschrieben. Nicht erörtert wird der weite Bereich der Fehlerbereinigung, da diese erst in der nächsten Projektphase durchgeführt werden wird.

## 5.1 Datenerfassung

Die Daten werden direkt am Bildschirm von den Erhebungsbögen in den Computer eingegeben. Numerische und Textdaten werden in einem Arbeitsgang erfaßt. Für die Datenerfassung wurde ein Maskenprogramm für den Siemens-EDT-Prozessor geschrieben. Dies Programm erfüllt folgende Funktionen:

### 1. Entlastung des Datentypisten von der Dateiverwaltung.

Der Datentypist benötigt dadurch nur geringe Kenntnisse des Betriebssystems. Er muß z. B. wissen, wie der Bildschirm aktiviert, der EDT-Prozessor geladen und das Erfassungsprogramm im EDT gespeichert wird. Die Datenabspeicherung, Zwischen- und Endsicherung, die Trennung in numerische und Textdatei, Umschalten auf Groß-Kleinschreibung, Drucken eines Protokolls der eingegebenen Werte und das Beenden des Eingabeprozesses wird über das Erfassungsprogramm gesteuert.

### 2. Prüfung auf gültige Werte.

Das Programm prüft, ob der jeweils eingegebene Wert in den gültigen Bereich einer Variablen fällt. Wird dieser Bereich unter- oder überschritten, so wird der Datentypist gezwungen, die Eingabe zu wiederholen.

### 3. Computergesteuerte Darbietung des Erhebungsbogens.

Einige Teile des numerischen Erhebungsbogens enthalten Informationen, die immer vorhanden sein müssen. Dazu gehören die Identifikationsnummer, die Nummer des Analytikers, das Geburtsdatum des Patienten, das Interviewdatum usw. An diesen Stellen wird der Datentypist automatisch durch den Erhebungsbogen geführt, indem die Kurzbezeichnung der jeweiligen Variablen auf dem Bildschirm erscheint, der Typist den entsprechenden Wert eingibt und danach die nächste Variable erscheint. Wegen der vielen fehlenden Informationen wäre diese automatische Führung bei allen Variablen aber zu zeitaufwendig. Für Bereiche, wo Sprünge zu erwarten sind, wurde deshalb eine halbautomatische Führung gewählt: Der Typist gibt die Spal-

# ZUMA

---

tennummer der jeweiligen Variablen an, das Programm gibt dann die entsprechende Variablenkennung auf den Bildschirm aus und der Typist setzt den Wert ein.

Im Textteil wird der Typist aufgefordert, die Abkürzung für das jeweilige Schlagwort einzugeben und danach den dazugehörigen Text. Das Programm ergänzt pro Eingabezeile die Identifikationsnummer und das angegebene Schlagwort.

#### 4. Sicherung der Eingabe.

Sobald die Eingabe des numerischen Teils eines Falles beendet ist, werden diese Informationen abgespeichert. Gleiches gilt für den Textteil. Damit ist sichergestellt, daß bei einem Systemzusammenbruch des Computers höchstens die Informationen des zuletzt bearbeiteten Falles verloren gehen. Welche Fälle tatsächlich erfaßt wurden, kann mit der Protokoll-Liste überprüft werden.

#### 5.2 Datenmanagement

Numerische und Textdaten werden in einer Datenbank verwaltet. Das Datenbankkonzept bietet nicht nur den Komfort, numerische und Textdaten zusammen zu verwalten, sondern es erhöht die Sicherheit vor unerwünschten Zugriffen auf die Daten und erlaubt fast beliebige Erweiterungen der Datenbasis. In dem hier verwendeten Datenbanksystem SIR ist es möglich, selbst einzelne Variable vor beliebigen Zugriffen zu schützen. In unserem Fall heißt das, daß z. B. künftig nur der Projektleiter auf die Identifikationsnummer des Patienten zugreifen kann.

Durch das Datenbankkonzept wird es auch leicht möglich sein, die Analyse-richtung vom Patient auf den Analytiker zu verlagern oder andere Informationen, wie sie z. B. bei der Inhaltsanalyse der Diagnosen entstehen werden, zu der Datenbasis hinzuzufügen. Mit SIR ist es möglich, neue Fälle auf einfachste Weise hinzuzufügen. Schließlich enthält die Datenbank nicht nur die Daten, sondern auch die Projekt-Dokumentation und alle notwendigen Datenmanipulationsprozeduren.

Nicht zu überschätzen ist die Eigenschaft von SIR, ein transportables System zu sein. Diese Eigenschaft besteht darin, daß die Datenbank nach ihrer Fertigstellung relativ einfach auf den Computer des SFI übertragen werden kann und dann dort allen Mitarbeitern des Instituts zur Verfügung steht.

## 6. Auswertungsbeispiele

In der Regel werden die Analysen des numerischen Teils mit SPSS, BMDP oder einem anderen Statistik-Paket durchgeführt werden. Der Übergang von der SIR-Datenbank in diese Pakete ist problemlos. Die Analyse des Textteils wird von TEXTPACK, einem Paket für die computerunterstützte Inhaltsanalyse, unterstützt. Die aus der Textvercodung resultierenden numerischen Werte werden in die Datenbank eingegeben und können an SPSS weitergegeben werden.

Alle Daten, die in den Beispielen genannt werden, sind "Spieldaten", die für Illustrationszwecke erzeugt wurden. Die "Ergebnisse" sind daher nicht identisch mit denen, die aus einer Analyse der Originaldaten resultieren. Deshalb werden in der Regel auch nur die Rohwerte und keine statistischen Maße angegeben.

### 6.1 Deskription der Population

Das erste Ziel des Projekts ist eine umfassende Deskription der Patientenspopulation. Diese Deskription erfordert - je nach Art der zugrundegelegten Information - verschiedene Verarbeitungsprozesse:

#### a) Direkte numerische Informationen der Erhebung:

Direkt numerisch erhobene Informationen, wie "Geschlecht des Patienten" oder "Kostenträger der Behandlung", können sofort in eine Deskription der Population eingehen. Für den Querschnitt werden solche Variablen ausgezählt (Tabelle 1) und für den Längsschnitt gegen das Interviewdatum (hier zu Gruppen von fünf Jahren zusammengefaßt) tabelliert (Tabelle 2).

# ZUMA

Tab. 1: Häufigkeitsauszählung Geschlecht des Patienten und Kostenträger der Behandlung

		ABSOLUTE HÄUFIGKEIT
MAENNLICH	1.00	221.00
WEIBLICH	2.00	250.00
	TOTAL	----- 471.00

		ABSOLUTE HÄUFIGKEIT
KOSTENTRAEGER-NICHTV	1.00	2.00
KOSTENTRAEGER-PRIVAT	2.00	85.00
KOSTENTRAEGER-RVO	3.00	73.00
KOSTENTRAEGER-ERSATZ	4.00	248.00
KOSTENTRAEGER-ANDERE	5.00	8.00
KOSTENTRAEGER-DSKV	6.00	5.00
KOSTENTRAEGER-ANDERE	8.00	3.00
	TOTAL	----- 424.00

STATISTICS.....  
GÜLTIGE BEOBACHTUNGEN 424  
FEHLENDE BEOBACHTUNGEN 45

Tab. 2: Geschlecht des Patienten kreuztabelliert mit Datum des Interviews (N = 270)

Geschlecht	Datum Interview			
	1960-1965	1966-1970	1971-1975	1976-1981
männlich	44	25	31	121
weiblich	50	21	30	149
Summe	94	46	61	270

b) Indirekte numerische Informationen der Erhebung:

Indirekte numerische Informationen sind numerische Informationen, die erst aus den direkten numerischen Informationen berechnet werden müssen. Beispiele sind das "Alter des Patienten zum Zeitpunkt des Interviews" und die "Position des Patienten in der Geschwisterreihe". Die Ausgangsinformationen für die Altersangabe sind das Geburtsdatum des Patienten und das Datum des Interviews (Tabelle 3), die Variable "Alter" wird dann durch die Subtraktion "Interviewdatum minus Geburtsdatum" erzeugt.

Tab. 3: Häufigkeitsauszählung der Variablen ALTER in 5-Jahresgruppen

Alter	Häufigkeit
17-20	16
21-25	87
26-30	130
31-35	122
36-40	59
41-45	31
46-50	11
51-75	14

Ähnlich kann die Position des Patienten in der Geschwisterreihe berechnet werden. In Abbildung 2 (Seite 51) ist der Teil des numerischen Bogens abgebildet, der sich auf die Geschwister des Patienten bezieht. Für bis zu sechs Geschwister werden die vollständigen Angaben erhoben. Die Gesamtzahl der Geschwister wird in der Rubrik "Anzahl" festgehalten. Um ganz sicherzugehen, daß die wichtige Angabe, ob der Patient Einzelkind ist, nicht durch Vercodungsfehler untergeht, wird diese Information direkt erfaßt. Hat der Patient nun Geschwister, so erhält man in der Regel auch die Informationen, ob diese älter oder jünger als der Patient sind. Aus dieser Angabe läßt sich nun die Geschwisterreihe leicht berechnen: Man zählt, wieviele Geschwister älter als der Patient sind und addiert "1" hinzu. D. h., sind bei 4 Geschwistern 2 älter als der Patient, dann ergibt sich "3" als die Position des Patienten. Mit Hilfe eines Filters (in SPSS: SELECT IF), kann man dann eine Häufigkeitsauszählung der Position solcher Patienten erzeugen, die keine Einzelkinder sind (Tabelle 4).

Tab. 4: Stellung des Patienten in der Geschwisterreihe

Position	Zahl der Fälle
1. Kind	196
2. Kind	121
3. Kind	42
4. Kind	22
5. Kind	4
6. Kind und mehr	8

In der Zahl des 1. Kindes sind 88 Einzelkinder enthalten.

Abb. 2: Numerischer Bogen - Geschwister des Patienten

23) GESCHWISTER DES PATIENTEN: (auch Halb-/Stiefgeschwister)		ANZAHL der Geschwister insgesamt:		keine (Einselkind)	
1. Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Altersdifferenz <input type="checkbox"/> <input type="checkbox"/> (in Jahren)	verstorben? vor Geburt Pat. <input type="checkbox"/> 1 nach Geburt Pat. <input type="checkbox"/> 2
2. Schwester	<input type="checkbox"/> 2	jünger	<input type="checkbox"/> 2		
KA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3		
Stief/Halb	<input type="checkbox"/> 1	KA	<input type="checkbox"/> 9	KA	<input type="checkbox"/> 9
1. Halbjahr = 82					
Alter Pat. bei Tod: <input type="checkbox"/> <input type="checkbox"/>					
2. Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Altersdifferenz <input type="checkbox"/> <input type="checkbox"/> (in Jahren)	verstorben? vor Geburt Pat. <input type="checkbox"/> 1 nach Geburt Pat. <input type="checkbox"/> 2
3. Schwester	<input type="checkbox"/> 2	jünger	<input type="checkbox"/> 2		
KA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3		
Stief/Halb	<input type="checkbox"/> 1	KA	<input type="checkbox"/> 9	KA	<input type="checkbox"/> 9
1. Halbjahr = 82					
Alter Pat. bei Tod: <input type="checkbox"/> <input type="checkbox"/>					
3. Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Altersdifferenz <input type="checkbox"/> <input type="checkbox"/> (in Jahren)	verstorben? vor Geburt Pat. <input type="checkbox"/> 1 nach Geburt Pat. <input type="checkbox"/> 2
4. Schwester	<input type="checkbox"/> 2	jünger	<input type="checkbox"/> 2		
KA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3		
Stief/Halb	<input type="checkbox"/> 1	KA	<input type="checkbox"/> 9	KA	<input type="checkbox"/> 9
1. Halbjahr = 82					
Alter Pat. bei Tod: <input type="checkbox"/> <input type="checkbox"/>					
4. Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Altersdifferenz <input type="checkbox"/> <input type="checkbox"/> (in Jahren)	verstorben? vor Geburt Pat. <input type="checkbox"/> 1 nach Geburt Pat. <input type="checkbox"/> 2
5. Schwester	<input type="checkbox"/> 2	jünger	<input type="checkbox"/> 2		
KA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3		
Stief/Halb	<input type="checkbox"/> 1	KA	<input type="checkbox"/> 9	KA	<input type="checkbox"/> 9
1. Halbjahr = 82					
Alter Pat. bei Tod: <input type="checkbox"/> <input type="checkbox"/>					
5. Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Altersdifferenz <input type="checkbox"/> <input type="checkbox"/> (in Jahren)	verstorben? vor Geburt Pat. <input type="checkbox"/> 1 nach Geburt Pat. <input type="checkbox"/> 2
6. Schwester	<input type="checkbox"/> 2	jünger	<input type="checkbox"/> 2		
KA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3		
Stief/Halb	<input type="checkbox"/> 1	KA	<input type="checkbox"/> 9	KA	<input type="checkbox"/> 9
1. Halbjahr = 82					
Alter Pat. bei Tod: <input type="checkbox"/> <input type="checkbox"/>					
6. Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Altersdifferenz <input type="checkbox"/> <input type="checkbox"/> (in Jahren)	verstorben? vor Geburt Pat. <input type="checkbox"/> 1 nach Geburt Pat. <input type="checkbox"/> 2
7. Schwester	<input type="checkbox"/> 2	jünger	<input type="checkbox"/> 2		
KA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3		
Stief/Halb	<input type="checkbox"/> 1	KA	<input type="checkbox"/> 9	KA	<input type="checkbox"/> 9
1. Halbjahr = 82					
Alter Pat. bei Tod: <input type="checkbox"/> <input type="checkbox"/>					

## c) Textinformationen:

Zu den Textinformationen zählen einige Merkmale demographischer Art wie die Angaben des Wohnorts und des Berufs. Für die Deskription erstellt man zuerst Häufigkeitslisten, indem man mit dem entsprechenden Programm von TEXTPACK gezielt auf die Schlagwörter WORT (Wohnort) und BPAT (Beruf-Patient) zugreift (vgl. Tabelle 5). Aufgrund der Häufigkeitsliste entwickelt man einen inhaltsanalytischen Diktionär, in dem einzelne Worte zu Kategorien zusammengefaßt sind. So könnte man für den Wohnort einen nach Entfernung des Wohnorts von Frankfurt klassifizierenden Diktionär entwickeln.

Tab. 5: Liste der Häufigkeiten der Themen Beruf des Patienten (BPAT) und Wohnort des Patienten (WORT) - Auszug -

Zahl der Nennungen	Wohnort
1	BENSHEIM-SCHWANHEIM
2	BRUCHKOEBEL
1	BUETTELBORN
1	DARMSTADT-ARHEILIGEN
1	EPPSTEIN
1	FLOERSHEIM-WEILBACH
254	FRANKFURT
2	FREIGERICHT
2	FRIEDRICHSDORF
2	FRANKFURT
1	GELNHAUSEN/
1	GERNSHEIM
1	GROSSKROTZENBURG
1	HAINBURG
1	HATTERSHEIM
2	HOFHEIM
1	INGELHEIM
1	KASSEL
3	KELKHEIM-FISCHBACH
1	KLEESTADT
1	KRIFTEL
1	LANGEN
1	MAJBACH
7	MAINZ
1	MESPELBRUNN
2	MOERFELDEN
1	MUENSTER-ALTHEIM

# ZUMA

---

## Nennungen Beruf

1	DIPLOM-CHEMIKERIN
1	DIPLOM-SOZIOLOGIN
1	DIPLOMBETRIEBSWIRT
1	DIPLOMBIOLOGIN
8	DIPLOMINGENIEUR
1	DIPLOMMATHEMATIKER
7	DIPLOMPAEDAGOGIN
4	DIPLOMPSYCHOLOGE
2	DIPLOMSOZIOLOGE
1	DIREKTIONSASSISTENTIN
1	DOKTORAND
1	DOLMETSCHERAUSBILDUNG
2	DOZENT
1	DROGISTIN
1	EINZELHANDELSKAUFMANN
1	ELEKTROMONTEURIN
1	ERGOTHERAPEUTIN
4	ERZIEHERIN
1	EVANGELISCHER
1	FACHLEHRER
1	FACHSCHUELER
1	FAHRLEHRER
1	FILIALLEITERIN

## Nennungen Beruf

1	DIPLOM-INGENIEUR
1	DIPLOM-VOLKSWIRT
1	DIPLOMBIBLIOTHEKARIN
1	DIPLOMGEOLOGIN
3	DIPLOMKAUFMANN
2	DIPLOMPAEDAGOGE
1	DIPLOMPHYSIKER
3	DIPLOMPSYCHOLOGIN
3	DIPLOMSOZIOLOGIN
2	DIREKTOR
1	DOKUMENTARIST
1	DOLMETSCHERIN
1	DR.RER.NAT.
1	EINZELHAENDLER
1	ELEKTROANLAGENINSTALLATEUR
1	EMPFANGSSEKRETAERIN
2	ERZIEHER
1	EXAMEN
3	FACHHOCHSCHULE
3	FACHLEHRERIN
1	FAHRER
1	FERNMELDEINGENIEUR
1	FILMEMACHER

Die Klassifikation der Berufe ist, wegen der sehr kurzen Angaben in den Akten, nur in grobe Raster möglich. Man könnte aber z. B. helfende und Lehrberufe gegenüber Dienstleistungen und Produktion kategorisieren. Der entsprechende Diktionär wird dann auf den Text angewandt. Das Ergebnis sind numerische Codes, die wie die Informationen des Quasi-Questionnaire tabelliert werden können.

## 6.2 Testen von Hypothesen

Ein weiteres Ziel des Projekts ist die Überprüfung bestimmter in der Literatur genannter Hypothesen. An zwei fiktiven Beispielen wird die Vorgehensweise bei solchen Tests dargestellt.

- a) Die Hypothese sei, daß Analytiker einem gegengeschlechtlichen Patienten häufiger Psychoanalyse empfehlen als einem gleichgeschlechtlichen. Diese Hypothese kann im ersten Anlauf mit einer Tabellierung der Merkmale "Psychoanalyse ja-nein", "Patient weibl.-männl." und "Analytiker weibl.-männl." getestet werden. Die notwendigen Informationen erzeugt man aus der numerischen Datei wie folgt: "Psychoanalyse ja-nein" ist in der nu-

# ZUMA

merischen Datei nur indirekt enthalten. Dazu muß Frage 18 des Quasi-Questionnaire "Ausbildungsanalyse"\* (siehe Anhang I) recodiert werden. Mit der Variablen "Ausbildungsanalyse" wird für jeden Patienten festgehalten, ob eine Ausbildungsanalyse möglich, eingeschränkt oder nicht möglich sei, oder ob diese Informationen nicht zutreffen, d. h. daß dem Patienten keine Psychoanalyse angeboten wurde. Der Recode lautet dann: "trifft nicht zu = 0, sonst = 1"; die Angabe des Geschlechts des Patienten ist direkt erhoben; das des Analytikers indirekt, indem die Kennungen der Analytiker implizit in zwei Bereiche geteilt sind (z. B. 1-n = weiblich und m-Ende = männlich). Durch einen entsprechenden Recode kann dann die Variable "Geschlecht des Analytikers" erzeugt werden und eine Tabelle für den Hypothesentest berechnet werden (Tabelle 6).

Tab. 6: Geschlecht des Therapeuten und des Patienten und empfohlene Therapieform

Geschlecht Therapeut	Geschlecht Patient			
	männlich		weiblich	
	Psychoanalyse	Keine Psychoanalyse	Psychoanalyse	Keine Psychoanalyse
männlich	65	33	78	63
weiblich	80	43	64	45

- b) Zu testen sei die Hypothese, daß die Diagnose nicht nur von Symptomen abhängt, sondern auch von der Alterskohorte des Analytikers. Dazu sind sicher umfangreiche Analysen notwendig, deren Ablauf hier nur skizziert werden kann. Auf der Seite der Symptome können z. B. Begriffscluster berechnet werden, d. h. Gruppen von Symptombezeichnungen, die häufig zusammen auftreten, wobei auch zu untersuchen ist, ob diese Cluster über die Zeit stabil blieben oder sich auf ganz bestimmte Altersgruppen bei den Patienten beziehen. Dann wäre die Variation dieser Cluster mit den Diagnosen und den Therapievor schlägen zu analysieren. Schließlich könnte dann die Alterskohorte des Analytikers als intervenierende Variable eingeführt werden und die Hypothese an sich überprüft werden.

\*"Ausbildungsanalyse" meint, daß ein noch nicht völlig ausgebildeter Analytiker (Kandidat) den Patienten unter Anleitung eines Mentors therapieren könnte.

## 6.3 Fallstudien

Die Datenbank wird es Analytikern am SFI erlauben, leichter als bisher Fallstudien durchzuführen. Das erste Hilfsmittel dazu werden standardisierte Krankenblätter sein. Es ist beabsichtigt, für die erfaßten Fälle eine Akte anzulegen, in der für jeden Fall die erfaßten Informationen in derselben Reihenfolge vorliegen werden (Abbildung 3). Daneben kann jeder Mitarbeiter für seine Zwecke durch Umsortieren von Teilen oder der ganzen Datenbasis seine ihm genehme "Quasibiographie" erzeugen.

Abb. 3: Auszug aus einer normierten Biographie

Ambulanznummer	9999
Therapeut	55
Wohnort	Frankfurt
Geboren am	2.12.1946
Datum des Interviews	6.6.1967
Konfession	keine Angabe
Überweisung	von sich aus / formale Überweisung
von	Hausarzt
Vorbehandlung	keine
.....	
PSYC	Beziehungsprobleme : hat Schwierigkeiten mit Partner
.....	
THEA	Psychoanalyse, Ausbildungsfall für fortgeschrittene Kandidaten
WEIT	keine Angabe

Ein weiteres Hilfsmittel für Fallstudien werden die Informationsrückgewinnungsfunktionen von SIR bzw. TEXTPACK sein (Informations-Retrieval). Hierzu zum Abschluß noch zwei Beispiele:

### a) Numerische Suchinformation.

Es sollen alle Fälle genauer untersucht werden, die von den Analytikern als Ausbildungsfälle ohne Einschränkungen bezeichnet wurden. Dazu wird ein Filter gesetzt, der alle Fälle auswählt, bei denen die Variable "Ausbildungsanalyse" den Wert "1" hat. Dann werden Teile der oder alle gesamten standardisierten Krankenblätter ausgedruckt; schließlich kann über die Identifikationsnummer auf die Originalakten zurückgegriffen

werden. Außerdem können zusätzliche Erhebungen, z. B. beim zuständigen Analytiker, angestellt werden.

## b) Nichtnumerische Suchinformation.

Gelegentlich empfiehlt die Ambulanzkonferenz eine andere Therapie als der Analytiker. In solchen Fällen unterscheiden sich die Informationen zu den Schlagwörtern "THERapievorschlAg Interview" und "THERapievorschlAg Ambulanzkonferenz". Die entsprechenden Fälle können per Programm oder per Hand ausgewählt werden und sind dann wie unter a) angegeben zu behandeln.

## Stand des Projekts

Die kumulierte Stichprobe ist vercodet (ca. 900 Fälle); damit ist dieser Teil des Projekts abgeschlossen. Aufgrund der ersten Auswertungen werden nun für die restlichen Fälle einige zentrale Merkmale erhoben, so daß in absehbarer Zeit alle Fälle des Archivs mindestens rudimentär erfaßt sind.

Die ersten Analysen bezogen sich nur auf einen Teil (ca. 2/3) der Stichprobe. Nunmehr sind alle Fälle, allerdings noch unbereinigt, erfaßt und in die Datenbank eingegeben. Gegenwärtig werden verschiedene inhaltsanalytische Diktionäre für die Grundausswertungen konstruiert; diese Diktionäre werden dann zusammen mit anderen Hilfsprogrammen in die Datenbank integriert. Den Abschluß der Arbeiten (Ende 1984) wird die Dokumentation der Grundausswertungen bilden.

Das Projekt wird von Otto Goldschmidt am Sigmund-Freud-Institut durchgeführt. Bei ZUMA betreut Peter Ph. Mohler, der auch diesen Beitrag verfaßte, das Projekt.

Anhang I (Seite 57-59) Quasi-Questionnaire

Anhang II (Seite 60) Liste der Kategorien Quasi-Text





Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Alteradifferenz	<input type="checkbox"/> 1	verstorben?	<input type="checkbox"/> 1	1. Halbjahr = 02
4. Schwester	<input type="checkbox"/> 2	junger	<input type="checkbox"/> 2			vor Geburt Pat.	<input type="checkbox"/> 1	Alter Pat
EA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3	(an Jahren)		nach Geburt Pat.	<input type="checkbox"/> 2	Bei Tod
Stief/Mutt	<input type="checkbox"/> 1	EA	<input type="checkbox"/> 9				EA	<input type="checkbox"/> 9
Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Alteradifferenz	<input type="checkbox"/> 1	verstorben?	<input type="checkbox"/> 1	1. Halbjahr = 02
5. Schwester	<input type="checkbox"/> 2	junger	<input type="checkbox"/> 2			vor Geburt Pat.	<input type="checkbox"/> 1	Alter Pat
EA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3	(an Jahren)		nach Geburt Pat.	<input type="checkbox"/> 2	Bei Tod
Stief/Mutt	<input type="checkbox"/> 1	EA	<input type="checkbox"/> 9				EA	<input type="checkbox"/> 9
Bruder	<input type="checkbox"/> 1	Alter	<input type="checkbox"/> 1	Alteradifferenz	<input type="checkbox"/> 1	verstorben?	<input type="checkbox"/> 1	1. Halbjahr = 02
6. Schwester	<input type="checkbox"/> 2	junger	<input type="checkbox"/> 2			vor Geburt Pat.	<input type="checkbox"/> 1	Alter Pat
EA	<input type="checkbox"/> 9	gleichaltrig (Zwilling/Drilling)	<input type="checkbox"/> 3	(an Jahren)		nach Geburt Pat.	<input type="checkbox"/> 2	Bei Tod
Stief/Mutt	<input type="checkbox"/> 1	EA	<input type="checkbox"/> 9				EA	<input type="checkbox"/> 9

24 WOHNSITUATION DER PATIENTEN.

EIGENE WOHNDUNG  C1

nicht im selben Haus wie Eltern  C1

im selben Haus wie Eltern  C1

UNTERMIET  C3

nicht im selben Haus wie Eltern  C3

im selben Haus wie Eltern, aber anderer Vermieter  C4

bei Eltern  C5

WOHNGEMEINSCHAFT  C6

nicht im selben Haus wie Eltern  C6

im selben Haus wie Eltern aber andere Wohnung  C7

im selben Haus wie Eltern mit Eltern  C8

WOHNHEIM  C9

ANDERE WOHNSITUATION (halboffer verboden)  C10

EA  99



# ZUMA

## Anhang II - Liste der Kategorien Quasi-Text

### PSYCHOANALYTISCHE AMBULANZ

Sigmund-Freud-Institut, Frankfurt/M.

ZUMA - Zentrum für Umfragen, Methoden und Analysen e.V., Mannheim 1 - B2,

### HALBOFFENER ERHEBUNGSBOGEN (Nov. 81)

<u>Kategorie</u>	<u>Stichwort</u>
Aufgewachsen bei	AUFW
Beruf Vater } in Kindheit (0 - 18)	BVAT
Beruf Mutter } in Kindheit (0 - 18)	BMUT
Beruf Patient } zum Zeitpunkt des Interviews	BPAT
Beruf Partner } zum Zeitpunkt des Interviews	BPAR
Eltern zusammen/getrennt von bis	EBEZ
Wohnungswechsel in Kindheit, unter dem gelitten wurde	WNEC
Psychische Störungen bzw. Erkrankungen	PSYC
Psychosomatische Störungen bzw. Erkrankungen	PSOM
Körperliche Erkrankungen	KORP
Unfälle	UNFA
Operationen	OPER
Ungewöhnliche schwere Ereignisse	TRAU
Leiden Vater	LEIV
Leiden Mutter	LEIM
Leiden Partner	LEIP
Leiden Geschwister	LEIG
Leiden restliche Familie	LEIF
Auslösende Situation	ASIT
Wohnsituation (soweit nicht numerisch erfaßbar)	WSIT
Wohnort	WORT
Interview-Diagnose:	
allgemein	DALI
Symptom	DSYI
Struktur	DSTI
Test-Diagnose:	
Struktur	DSTT
Ambulanzkonferenz-Diagnose bzw. abschließende Diagnose:	
allgemein	DALA
Symptom	DSYA
Struktur	DSTA
Therapievorschläge:	
im Interviewbericht	THEI
im Testbericht	THET
Abschlußvorschlag	THEA
Abwehrmechanismen	ABWE
Abbruch der Untersuchungen	ABBR
Weiterbehandlung	WEIT
Dramatische Reaktionen Patient im Interview	DREP

## Anhang 1

### Q U A L S H O P I

#### Programm

Zeit: 28.2. bis 2.3.1983

Ort: Universität Mannheim (Schloß)  
Raum 104 in der Mensa

#### Montag, 28.2.1983

- 10-13 Uhr Vorkurs - Einführung in die Grundbegriffe des Datenmanagements in der EDV (Mohler)
- 14-17 Uhr Begrüßung (Küchler)
- Datenmanagement bei qualitativen Erhebungsverfahren: ein Überblick (Mohler)
- Integrationsmöglichkeiten von Standardsoftware der Textverarbeitung mit kompatiblen FORTRAN-Programmen (Schupp)

#### Dienstag, 1.3.1983

- 10-13 Uhr Erfahrungen mit umfangreichen qualitativen Dateien Probleme deren Analyse und Präsentation (Kreibich)
- Planung und Entwicklung eines Programmsystems zur EDV-gestützten qualitativen und quantitativen Analyse von Verbaldaten (Paulus)
- 14-17 Uhr SIR 2 - ein komfortabler Zettelkasten für die Textanalyse? (Höllbacher)
- Ungenauigkeit und Effizienz. Die Informationsstruktur historischen Quellenmaterials und ihre Bearbeitung mit dem datenbankorientierten Programmsystem CLIO (Thaller)

Mittwoch, 2.3.1983

10-13 Uhr Probleme der Texterfassung (Mohler)  
KDM-Omnifont: Die Maschine, die liest (Saake)

14-16 Uhr Was tun? (tutti unter Leitung von Kuchler)