Inhalt

Vorbemerkungen

von Ulrich Engel

Der vorliegende zweite Forschungsbericht gibt in Teil I und II Rechenschaft von der maschinellen Sprachbearbeitung im Institut für deutsche Sprache. Von Anfang an hat dieser Forschungsbereich im Rahmen der gesamten Institutsarbeit eine große Rolle gespielt. In der Außenstelle Bonn begann die maschinelle Textaufnahme Anfang 1965, in der Mannheimer Arbeitsstelle im Sommer 1965. Es ist nun der Zeitpunkt gekommen, das bisher Erreichte darzulegen und weitere Pläne und Möglichkeiten zu umreißen.

Die bisherige Arbeit brachte viel Mühe für die unmittelbar Beteiligten, auch mancherlei Enttäuschungen; wir können jetzt immerhin sagen, daß der zurückgelegte Weg mit soliden Erfahrungen gepflastert ist. Diese Erfahrungen sind nützlich auch insofern, als sie uns vor ungerechtfertigten Hoffnungen bewahren. Nirgends sind so viele Blütenträume zerstört worden wie auf dem Gebiet der maschinellen Sprachbearbeitung. Ehe nun neue Aufgaben in Angriff genommen werden sollen, ist es vor allem wichtig zu wissen, was wir n i c h t erreichen können.

Uns ist immer wieder die Frage gestellt worden, ob maschinelle Sprachbearbeitung "rentabel" sei. Es geht sicher nicht an, eine solche Fragestellung einfach mit dem Hinweis abzulehnen, Rentabilitätsberechnungen gehörten gar nicht in den Bereich der Forschung. Wissenschaft kostet Geld, und wer sie finanziert, hat ein Recht darauf zu erfahren, ob er sein Geld vernünftig, zweckentsprechend, wirksam angelegt hat.

Wir beginnen mit dem Bekannten. Wo überlieferte Arbeitsgänge, die bisher "manuell" ausgeführt wurden, nun von der Maschine übernommen werden, läßt sich die Frage, wer rationeller arbeitet, gar nicht abweisen. Die Ant-

worten werden unterschiedlich ausfallen müssen. Sicher sind Zähl- und Sortiervorgänge sowie komplizierte Rechenoperationen vom Computer wesentlich schneller und auch wesentlich preiswerter durchzuführen als von einem einzelnen Forscher oder auch von Forscherteams. Und auch die Codierung grammatischer Merkmale auf peripheren Datenträgern (Lochkarten, Lochstreifen) läßt sich einfacher, zudem exakter durchführen als auf die herkömmliche Art (Schreiben auf Karteikarten). Wenn aber eine unmittelbare maschinelle Bearbeitung des "Klartextes" vorgesehen ist, muß dieser Text erst für die Maschine lesbar gemacht, d.h. in bestimmter Form und Anordnung auf Magnetbänder übertragen werden. Erhebliche Schwierigkeiten bereiten vor allem die zahlreichen Korrekturen, die notwendig sind, bis der Text in brauchbare Form gebracht ist. Fast alle unsere ursprünglichen Zeitplanungen haben sich als zu optimistisch erwiesen, weil sie das tatsächliche Ausmaß der erforderlichen Korrekturgänge unterschätzt hatten. Allerdings darf man nicht übersehen, daß Texte, einmal so aufbereitet, für theoretisch unbegrenzte Zeit und für sehr viele und verschieden geartete maschinelle Auswertungen zur Verfügung stehen. Da für alle späteren Bearbeitungen die zeit- und kostenraubenden Schreib- und Korrekturvorgänge für den Klartext ja nicht mehr vorgenommen werden müssen, wird sich das Verhältnis von Aufwand und Erfolg sehr bald günstiger gestalten. So kann man zusammenfassend sagen, daß Untersuchungen begrenzten Ausmaßes, zumal wenn Klartexte zugrunde gelegt werden, nicht immer den Einsatz der Maschine rechtfertigen; oft ist dann die herkömmliche "manuelle" Methode wirtschaftlicher. Aber im allgemeinen wird die elektronische Speicherung sprachlicher Texte ohnehin im Hinblick auf zahlreiche Untersuchungsmöglichkeiten erfolgen. Nur wer in sehr kurzen Zeiträumen planen muß, kann maschinelle Sprachbearbeitung "unrentabel" finden.

Hinzu kommt aber außerdem, daß die Aufgabe der maschinellen Sprachbearbeitung keineswegs nur in der Übernahme bisher "manuell" durchgeführter, aber mechanisierbarer Arbeitsvorgänge besteht. Vielmehr kann die elektronische Rechenmaschine Probleme lösen, die mit konventionellen Mitteln gar nicht zu lösen wären, und sie kann auch zu völlig neuen Fragestellungen führen. Die

Ausweitung des Gesamtbereichs der Linguistik im vergangenen Jahrzehnt, wie sie sich an der zunehmenden Bedeutung der Nachbarwissenschaften (formale Logik, Informationstheorie, Mengentheorie seien als Beispiele genannt) ablesen läßt, hat Probleme aufgeworfen, für deren Lösung die Hilfe des Computers unentbehrlich geworden ist. Natürlich verursacht die Lösung dieser Probleme Kosten, die in früherer Zeit einfach deshalb nicht entstanden sind, weil die Probleme noch gar nicht gestellt waren; nur aus wirtschaftlichen Gründen auf die Lösung von Problemen verzichten zu wollen, würde aber einen irreparablen wissenschaftlichen Rückschritt bedeuten.

Zur maschinellen Sprachbearbeitung in Mannheim (Teil I)

Die Datenerfassung auf Lochstreifen, in besonderen Fällen auch auf Lochkarten, erfolgt in den Räumen des Instituts in Mannheim, wo die unter 2.1.(1) aufgeführten Geräte aufgestellt sind. Die Übertragung auf Magnetbänder und die weitere Verarbeitung der Texte geschieht im Deutschen Rechenzentrum in Darmstadt. Die in Darmstadt anfallenden Arbeiten wurden in der ersten Zeit von Gerhard Stickel, einem Mitarbeiter des Deutschen Rechenzentrums, durchgeführt. Nach seinem Weggang 1966 übernahm Ingeborg Z i n t, M.A., die Fortführung dieser Arbeiten. Sie erwiesen sich mit zunehmender Menge der anfallenden Texte als so umfangreich, daß sie die Arbeitskraft eines einzelnen Mitarbeiters weit überschritten. Deshalb wurde im Herbst 1967 als weiterer Mitarbeiter für die maschinelle Sprachbearbeitung Paul W o l f a n g e l, M.A., angestellt. Aber die Forschungsgruppe ist angesichts der zu lösenden Aufgaben auch mit zwei wissenschaftlichen Mitarbeitern völlig unterbesetzt. Wenn die Textaufbereitung beschleunigt und zugleich weitere Aufgaben in Angriff genommen werden sollen, wie es unseren Absichten entspricht, müßten mindestens fünf volle wissenschaftliche Kräfte zur Verfügung stehen.

Angesichts der aufgezeigten Schwierigkeiten und der unzulänglichen personellen Verhältnisse kann das bisher Erreichte als zufriedenstellend bezeichnet werden.

Die unter 1.1. aufgeführten Quellen ergeben zusammen ein Corpus deutscher Gegenwartssprache, das rund 1,5 Millionen Wörter umfaßt. Von zehn dieser Texte liegen zur Zeit (Ende 1968) die unter 3.1. beschriebenen Register vor. Es ist anzunehmen, daß sich diese Zahl in nächster Zeit erhöhen wird, da der größte Teil der übrigen Texte schon mehrfach korrigiert wurde. Lediglich an der "Bildzeitung" (7 Monate in Auswahl) wird noch geschrieben. Im abgelaufenen Jahr 1968 kam die Sprachbearbeitung nur langsam vorwärts, weil der Rechenbetrieb im Deutschen Rechenzentrum in Darmstadt infolge länger dauernder Umbauarbeiten immer wieder gestört war. Bemühungen, einen Teil der Rechenarbeiten, besonders die zeitraubenden Korrekturen, auf anderen und günstiger gelegenen Rechenanlagen durchführen zu lassen, haben bis jetzt nicht zum Erfolg geführt. Es ist bei alldem hervorzuheben, daß die Zusammenarbeit mit den Mitarbeitern des Deutschen Rechenzentrums immer ausgezeichnet war. Der Direktion und vielen Mitarbeitern des Deutschen Rechenzentrums, namentlich Friedhelm S c h u l t e - T i g g e s als Leiter der Abteilung "Nichtnumerik", haben wir an dieser Stelle unseren herzlichen Dank auszusprechen. Für wertvollen Rat und förderliche Hilfe haben wir außerdem zu danken Professor Dr. Hans E g g e r s und seinen Mitarbeitern von der Universität Saarbrücken sowie dem Institut für Phonetik und Kommunikationsforschung in Bonn, besonders dessen Direktor, Professor Dr. Gerold U n g e h e u e r, und Dr. Dieter K r a l l m a n n. Die Umwandlung unserer Wortformregister in Grundformregister, von der wir uns besonderen Nutzen versprechen, wurde bisher ausschließlich von Dieter Krallmann durchgeführt.

Die unter 1.1. wiedergegebene Quellenliste kann nach dem Stande der Forschung keinen Anspruch darauf erheben, die Zusammensetzung des deutschen Gegenwartsschrifttums zuverlässig abzubilden. Alle daraus gewonnenen Erkenntnisse können also zunächst nur für diese Quellen gelten; eine Übertragung der Ergebnisse auf die deutsche Gegenwartsprache in ihrer geschriebenen Form ist nicht ohne weiteres zulässig. Ein für das deutsche Gegenwartsschrifttum r e p r ä s e n t a t i v e s Corpus könnte diese Texte übernehmen. Erweiterungen wären zweifellos nötig, vor allem (aber nicht nur) durch die sogenannte Gebrauchsliteratur.

Andererseits würde sich ein Teil der Texte voraussichtlich als unnötig umfangreich erweisen; statt des ganzen Werkes könnte jeweils ein Ausschnitt genügen. Letzten Endes ist das Problem wahrscheinlich überhaupt nicht vollständig lösbar. Repräsentativität einer Textauswahl baut sich auf sehr vielen und teilweise divergenten Merkmalen auf: zu berücksichtigen sind literarische Gattung und Thema, besondere Intention des Autors, Aufnahme und Bewertung durch die Leserschaft (was durchaus nicht dasselbe ist), das spezifische linguistische Untersuchungsobjekt u.a. Besonders der letztgenannte Gesichtspunkt ist bei Bemühungen um repräsentative Auswahlcorpora zu wenig beachtet worden. Für statistische Wortschatzerhebungen, auch für Ermittlungen über allgemeine Satzstrukturen (Satzbaupläne) genügen im allgemeinen kleine Corpora; bei Erweiterung ändern sich die relativen Werte nur noch ganz geringfügig. Geht es dagegen um seltene Erscheinungen wie das Passiv oder gewisse Merkmalskombinationen, so müßte ein Corpus, um "repräsentativ" zu sein, so umfangreich angelegt werden, daß eine Bearbeitung in angemessener Frist nicht mehr möglich wäre. Repräsentativ wird das Auswahlcorpus immer nur in dieser oder jener Hinsicht sein; dem Ideal eines schlechthin repräsentativen Corpus kann man sich nur möglichst weitgehend anzunähern versuchen. Bemühungen dazu sind im Gange.

Bisher wurden in Mannheim nur geschriebene Texte bearbeitet. Ein Corpus g e s p r o c h e n e r Texte, das auf einen endgültigen Umfang von mindestens 600.000 Wörtern angelegt ist, wird zur Zeit unter Leitung von Professor Dr. Hugo S t e g e r (Kiel, jetzt Freiburg i.Br.) und auf Grund einer von seinen Mitarbeitern entworfenen Typologie in einer Außenstelle des Instituts zusammengestellt; die Frage der repräsentativen Auswahl wird dabei von Anfang an stärker berücksichtigt. Die maschinelle Aufbereitung dieser Texte ist angelaufen.

Natürlich bestehen über die unter 3.2. beschriebenen Auswertungsmöglichkeiten hinaus weitere Pläne. Da grammatische Probleme der deutschen Gegenwartssprache im Mittelpunkt der gegenwärtigen Arbeiten des Instituts stehen, wird angestrebt, die Maschine im besonderen für diese Arbeiten nutzbar zu machen. Das von

Dr. Alex S t r ö b l unter 4. beschriebene Verfahren ist ein möglicher Weg. Die automatische Textanalyse ist allenfalls ein Fernziel, dem wir uns nur in kleinen Schritten annähern können. Ausgangspunkt dafür könnte ein Valenzwörterbuch sein, das zur Zeit im Institut für deutsche Sprache erstellt wird. Immer werden dabei, darüber sind wir uns im klaren, maschinelle und "manuelle" Analyse Hand in Hand gehen, schon deshalb, weil ohne ein wenn auch minimales Lexikon keine maschinelle Analyse möglich ist, und weil dieses Lexikon ständig von Hand ergänzt werden muß. Wertvolle Hinweise zum syntaktischen Wörterbuch verdanken wir wieder Professor Eggers - Saarbrücken. Auch für ein von Dr. phil.habil. Paul G r e b e vorgeschlagenes syntagmatisches Arbeitsvorhaben, das die Zuordnungsmöglichkeiten des deutschen Wortschatzes erfassen und systematisieren soll und das voraussichtlich im kommenden Jahr anlaufen wird, erhoffen wir uns die Hilfe der Maschine.

Es empfiehlt sich in diesem Zusammenhang, zu betonen, daß die maschinelle Sprachübersetzung nicht zu den Zielen des Instituts für deutsche Sprache gehört.

Zur maschinellen Sprachbearbeitung in Bonn (Teil II)

Der Bericht von Dr. Manfred H e l l m a n n von der unter der Leitung von Professor Dr. Hugo M o s e r stehenden Außenstelle Bonn beschränkt sich nicht auf die maschinelle Aufbereitung west- und ostdeutscher Zeitungstexte für Wortschatzuntersuchungen, sondern beschäftigt sich darüber hinaus mit der Corpusgewinnung und der Ermittlung repräsentativer Textauswahl überhaupt, Problemen, die auch in Mannheim seit langem diskutiert werden, und deren wissenschaftliche Erörterung einem weiteren Forschungsbericht vorbehalten bleiben soll. Der Bonner Bericht ist deshalb verhältnismäßig umfangreich geworden. Er hat dafür den Vorteil, den wichtigsten Bereich der derzeitigen Forschungen der Bonner Außenstelle vollständig und in aller Ausführlichkeit zu schildern.

Daneben werden in der Außenstelle Bonn kleinere Spezialuntersuchungen angefertigt, und die Veröffentlichungen zum Thema "Sprache im geteilten Deutschland" werden möglichst vollständig gesammelt. Seit Oktober 1967 arbeitet Dr. Arne Schubert an einer Bibliographie raisonnée der osteuropäischen, besonders der sowjetrussischen Germanistik und Linguistik. Schließlich unterstützte und unterstützt die Außenstelle eine Reihe wissenschaftlicher Arbeitsvorhaben durch Bereitstellung maschinell sortierten Materials.

An den vorbereitenden Arbeiten zur Dokumentation von Zeitungstexten war anfänglich auch die damalige wissenschaftliche Mitarbeiterin Dr. Inge Kraft beteiligt. Seit Februar 1965 lag diese Aufgabe allein in den Händen von Manfred Hellmann, dem stellvertretenden Leiter der Außenstelle Bonn. Für wertvolle Hilfe ist das Institut wiederum dem Institut für Phonetik und Kommunikationsforschung in der Universität Bonn und hier besonders Professor Dr. Gerold Ungeheuer und seinem Mitarbeiter Dr. Dieter Krallmann zu Dank verpflichtet.

Der bei den Bonner Arbeiten eingeschlagene Weg, aus Zeitungsjahrgängen eine Auswahl aufzunehmen, die als Modell des betreffenden Zeitungsjahrganges gelten kann, war in dieser Form neu. Weder die Zeitungswissenschaft noch die linguistisch orientierte Statistik noch die maschinell orientierte Linguistik haben für ein solches Verfahren bisher Vorbilder geschaffen. Insofern ist der vorliegende Bericht notwendigerweise zuerst Problembericht und sollte auch dort, wo das gewählte Verfahren konkret dargestellt wird, vor allem als Voschlag und als Diskussionsgrundlage verstanden werden.

Die Bonner Arbeiten zur Dokumentation von Zeitungstexten dienen zwar linguistischen Zwecken, berühren jedoch auch viele andere Gebiete. In besonderem Maße war die Außenstelle daher auf vielseitige Hilfe angewiesen. Sie dankt

allen Helfern, ohne jeden im einzelnen nennen zu können. Namentlich gedankt sei außer dem schon erwähnten Bonner Institut für Phonetik und Kommunikationsforschung Herrn Schütz, Bonn, dem Institut für Publizistik der Freien Universität Berlin, Herrn Dr. Leimbach und Frau Dr. Schuster für die Beschaffung der Zeitungstexte und Photokopien, Professor Dr. Unger und Professor Dr. Peschl, den Direktoren des Rheinisch-Westfälischen Instituts für instrumentelle Mathematik in Bonn, für die Erlaubnis, die Rechenanlage IBM 1410/7090 zu benutzen, ebenso vielen Mitarbeitern des Instituts, im besonderen Dr. habil. Krückeberg.

Über die Signifikanz bei Auswahltexten (Teil III)

Dieser Teil des Forschungsberichts befaßt sich mit einer Frage, die sich im Zusammenhang mit den linguistischen Arbeiten im Institut für deutsche Sprache mit zunehmender Dringlichkeit stellt. Man weiß, daß auch eine sehr sorgfältig zusammengestellte Auswahl die Eigenart und die statistischen Verhältnisse der Gesamtmenge, der sie entnommen wurde, nie hundertprozentig wiederspiegelt. Wesentlich ist dabei, in welchem Maße eine Auswahl der Gesamtmenge entspricht und wie der Grad der Entsprechung meßbar und beschreibbar gemacht werden kann. Auch dafür gibt es bisher keine Vorbilder. Das in Teil III geschilderte Verfahren wurde von cand.phil. Günther Billmeier - er ist seit 1965 studentischer Mitarbeiter der Außenstelle Bonn und war vorübergehend fest angestellter Programmierer - im Sommer 1967 an einem Teil der in der Außenstelle verfügbaren Textmengen entwickelt. Ob es in dieser Form brauchbar ist, kann erst eine Erprobung an mehreren anderen und größeren Textmengen erweisen. Es geht also hier wiederum mehr darum, das Verfahren als solches zur Diskussion zu stellen, weniger um die errechneten Ergebnisse. Eine solche kritische Diskussion erscheint auch deshalb wünschenswert und notwendig, weil Verfahren zur Gewinnung einer qualitativ und quantitativ zureichenden Textauswahl auch über den Bereich des Instituts für deutsche Sprache hinaus von allgemein linguistischem Interesse sind.

I. Maschinelle Sprachbearbeitung des Instituts für deutsche Sprache in Mannheim

von Ingeborg Zint

ÜBERSICHT :

0. Einleitung.

Im Rahmen der Institutsarbeit zur Dokumentation der Gegenwartssprache kommt der maschinellen Sprachbearbeitung erhebliche Bedeutung zu. Der folgende Bericht gibt im wesentlichen einen Überblick über bisher Erreichtes, vermeidet Diskussionen über Grundsätzliches und enthält sich programmiertechnischer Einzelheiten, die eher Gegenstand einer institutsinternen Arbeitsanleitung als eines Arbeitsberichts sein können. Der Leser soll in die Lage versetzt werden, sich eine Vorstellung von dem zur Zeit lieferbaren Output zu machen, der Gegenstand weiterer linguistischer Bearbeitung sein kann. Wege und Umwege, die zu diesem Output führen, werden nur insoweit beschrieben, als sie sich inzwischen stabilisiert haben (vgl. Punkt 2), und in dem Maße, wie sie für den mit der Programmierung nicht vertrauten Linguisten von Interesse sind. Aus diesem Grunde wird beispielsweise auf eine detaillierte Programmbeschreibung verzichtet. Ebenso widerspricht es der erwähnten Intention dieses Berichtes, Auswahl und Zusammensetzung der verarbeiteten Textmengen zu diskutieren.

Das unter Punkt 4 vorgestellte Verfahren "Parallelcodierung" geht auf eine Anregung von Herrn A. Ströbl zurück und wurde auch von ihm für die vorliegenden Ausführungen beschrieben.

1. Textbibliothek.

Hinter dem Etikett "Textbibliothek" verbirgt sich ein nach bestimmten Kriterien ausgewähltes Corpus geschriebener deutscher Gegenwartssprache - es wird etwa ein Zeitraum von zwanzig Jahren erfaßt -, das für die Auswertung durch elektronische Rechenanlagen präpariert und archiviert wird. Das gegenwärtig bearbeitete Material umfaßt Werke aus den verschiedensten Bereichen: Dichtung, Trivialliteratur, wissenschaftliche und populärwissenschaftliche Literatur, Memoiren, Zeitungen und Zeitschriften.

Es sei an dieser Stelle darauf hingewiesen, daß gleichlaufende Arbeiten zur Erstellung eines Corpus für die gesprochene Sprache im Gange sind. Aufnahme und Aufbereitung für den maschinellen Zugriff werden - wegen der geforderten Vergleichbarkeit von gesprochener und geschriebener Sprache - vorbehaltlich notwendiger Modifikationen ähnlich gehandhabt, wie es im vorliegenden Bericht für die geschriebene Sprache beschrieben wird.

1. 1. Liste der Texte.

Jeweils in der Kopfleiste befinden sich die Siglen für die maschinelle Textkennzeichnung:

a) Dichtung

LBT	Bergengruen, Werner, Das Tempelchen, 1950
LBC	Böll, Heinrich, Ansichten eines Clowns, 1963
LFH	Frisch, Max, Homo Faber, 1965
LGB	Grass, Günther, Die Blechtrommel, 1964
LMB	Mann, Thomas, Die Betrogene, 1954
LSO	Strittmatter, Erwin, Ole Bienkopp, 1963
LJA	Johnson, Uwe, Das dritte Buch über Achim, 1961

b) Trivialliteratur

TJM	Jung, Else, Die Magd vom Zellerhof, o.J.
TPM	Pinkwart, Heinz, Mord ist schlecht für hohen Blutdruck, 1963
TSH	Stauffen, Pia, Solange dein Herz schlägt, o.J.

c) Wissenschaftliche und populärwissenschaftliche Literatur

WBO	Bamm, Peter, Ex ovo, 1963
WBM	Bollnow, Otto, Maß und Vermessenheit des Menschen, 1962
WGW	Gail, Otto Willi, Weltraumfahrt, 1958
WGS	Grzimek, Bernhard, Serengeti darf nicht sterben, 1964
WHK	Heimpel, Hermann, Kapitulation vor der Geschichte, 1960
WHN	Heisenberg, Werner, Das Naturbild der heutigen Physik, 1963
WJA	Jaspers, Karl, Die Atombombe und die Zukunft des Menschen, 6
WJZ	Jungk, Robert, Die Zukunft hat schon begonnen, 1952
WPE	Pörtner, Rudolf, Die Erben Roms, 1965
WSP	Staiger, Emil, Grundbegriffe der Poetik, 1963
WUB	Ullrich, Fritz, Wehr dich Bürger!, 1960

d) Memoiren

MHE	Heuß, Theodor, Erinnerungen 1905-1933, 1964

e) Zeitungen und Zeitschriften

ZFA	"Frankfurter Allgemeine Zeitung", 1 Monat (1966)
ZWE	"Die Welt", 3 Monate (1965-66)
ZBW	"Das Bild der Wissenschaft" (Heft 1, 2 und 3, 1967)
ZSG	"Studium Generale" (Heft 12, 1966)
ZUR	"Urania" (Heft 11, 1966, Heft 1, 1967)
ZBZ	"Bildzeitung", 7 Monate (1967)

2. Textaufbereitung.

Bei der maschinellen Textaufbereitung stellen sich zunächst Fragen nach dem geeigneten Datenträger, der angemessenen Programmierung und dem optimalen Verfahren. Lösungen, die sich hier anboten, waren zum Teil vorgezeichnet durch die personellen und sachlichen Gegebenheiten zu dem Zeitpunkt, als das Institut diese Arbeit in Angriff nahm. Der einmal eingeschlagene Weg wurde im wesentlichen beibehalten, jedoch immer dann modifiziert, wenn es nötig und gleichzeitig auch möglich war.

2.1. Datenträger.

Als Datenträger für die Aufbereitung dienen 8-Kanal- und 5-Kanal-Lochstreifen und Lochkarten, für die Auswertung Magnetbänder. Das Institut in Mannheim verfügt über zwei Lochstreifenschreiber vom Typ "Supertyper" ohne Lesegerät, seit kurzer Zeit über einen weiteren mit angeschlossenem Lesegerät, sowie über einen Schreiblocher IBM 026. Etwa 70 Magnetbänder sind in Betrieb, z.T. als sogenannte Archivbänder (d.h. sie dienen der Speicherung von Daten und Ergebnissen), z.T. als Arbeitsbänder.

2.2. Programmierung.

Die benutzten Programme sind in FORTRAN II und FAP für die IBM-Rechenanlagen 1011, 1401 und 7094, wie sie im DRZ Darmstadt vorhanden sind, geschrieben. Es existieren einige Programme in FORTRAN IV und MAP (vgl. Punkt 4). Bei allen Programmen wurden die Hilfsmittel benutzt, die die im DRZ vorhandene Programmbibliothek für die Nichtnumerik bietet. - Eine detaillierte Beschreibung aller benutzten sowie der im Institut erstellten Programme mit genauen Angaben über Leistung und Verwendbarkeit hinsichtlich ihrer Organisation liegt gesondert vor.

An dieser Stelle mag eine kurze Charakterisierung der Programme UMCODIERUNG, CORRECTURE und KORREKTUR genügen. Diese am häufigsten verwendeten Standardprogramme sind als sogenannte Chain-Jobs organisiert, deren einzelne Links je nach Zielsetzung ausgetauscht bzw. kombiniert werden können.

UMCODIERUNG besteht aus:

a) STRCOR - Die Zeichen der Lochstreifenaufnahme werden umcodiert, z.B. die Umschaltzeichen von Groß- auf Kleinschreibung und umgekehrt.

b) UMLAUT - Ein Umschaltzeichen für Großschreibung erscheint als Stern hinter dem gefundenen Nomen. Die Umlaute ö, ä, ü erscheinen als oe, ae, ue, gleiches gilt für ß → ss.
Die geblockten Bandsätze werden in ungeblockte Sätze der Länge 21 Speicherwörter umgesetzt.

Die Ausgabe liefert einen Text mit fortlaufender Zeilennumerierung.
Bei dem Programm CORRECTURE ist den beiden oben beschriebenen Links noch ein Link CORREC vorgeschaltet, das es erlaubt, in ein und demselben Lauf den umzucodierenden Text nach Eingabe zweier Datenbänder (ein Originalband mit Zeilennumerierung und ein Band mit Korrekturzeilen) herzustellen. Ausgabe wie bei UMCODIERUNG.

KORREKTUR besteht aus den drei Links von CORRECTURE und einem weiteren Link SAETZE. Die Ausgabe liefert

a) einen Text mit Zeilennumerierung und
b) einen in Sätze zerlegten Text.

2.3. Verfahren bei der Aufbereitung.

Der Arbeitsablauf der Textaufbereitung stellt sich wie folgt dar: die unter 1.1. aufgeführten Texte werden unter Berücksichtigung der unten verzeichneten Ablochvorschriften mit Hilfe der Supertyper auf 8-Kanal-Lochstreifen geschrieben. Anschließend wird der Inhalt der Lochstreifen auf Magnetband übertragen, wobei der Text automatisch durchlaufend zeilenweise (auch Leerzeilen) in Zehnerschritten numeriert wird. Diese Numerierung hat den Sinn, nachfolgende Korrekturen zu erleichtern. Es ergibt sich dadurch die Möglichkeit, für anhand des Maschinenausdrucks als fehlerhaft erkannte Zeilen einen neuen Lochstreifen zu schreiben, diesen auf ein zweites Magnetband zu übertragen und den Inhalt beider Bänder per Programm derart zu mischen, daß jeweils eine falsche Zeile durch eine richtige ersetzt wird. Es ist möglich und meistens auch nötig, dieses Korrekturverfahren mehrfach zu wiederholen.

2.4. Schreibkonventionen.

Maschinelle Textauswertung (hier Verzetteln des Materials im weitesten Sinn) setzt automatisches Identifizieren des Textes in all seinen Teilen voraus. Dazu einige erläuternde Bemerkungen: Ein in den Lochstreifencode umgesetzter Text läßt sich beschreiben als ein lineares Kontinuum von einzelnen und in Kombination vorkommenden alphamerischen Zeichen, die jeweils durch eine Leerstelle voneinander getrennt sind. Der Zeichenvorrat umfaßt z.B. Buchstaben, Satzzeichen, Ziffern, Operationszeichen und Sonderzeichen wie $, * usw.. Jede Zeichengruppe - wobei eine Gruppe auch aus nur einem Zeichen bestehen kann - zwischen Leerstellen gilt als Wort. Der Wechsel von der gegliederten "zweidimensionalen" Abbildung eines Textes im Buch zu der "eindimensionalen" Darstellung auf dem Lochstreifen erfordert gewisse angemessene Zusatzcodierungen. Im Buch ist eine beliebige Ziffernfolge z.B. dann

als Seitenzahl definiert, wenn sie an einer ganz bestimmten Stelle rechts, links, in der Mitte über oder unter dem Text steht. Beim Lochstreifen ist das Merkmal "Stellung" vergleichsweise nicht realisierbar; es wird ersetzt durch "Kombination mit einem bestimmten Zeichen in festgelegter Form" - bei der Seitenzahl: 6mal "s" mit direkt anschließender sechsstelliger, rechtsbündiger Ziffernfolge -, wobei die Form der Kombination aus programmtechnischen Gründen so und nicht anders gewählt wurde. Derartige Zusatzcodierungen werden dem Streifentext außer bei Seitenangaben noch bei Buchtiteln (ttttt "Buchtitel") und Überschriften (uuuuuu "Überschrift") hinzugefügt.
Unter dem Aspekt von Wortschatzuntersuchungen und im Hinblick auf statistische Auswertungen hat es sich als nötig erwiesen, auch die folgenden Textteile eindeutig formal zu kennzeichnen:

1. fremdsprachige Textteile;
2. Unter- oder Beischriften zu Bildern, Tabellen u.ä.;
3. Quellenangaben, wie z.B. "Bonn, den 28.1.64 (dpa)";
4. Verfasserangaben in Zeitungsartikeln;
5. Zitate;
6. mathematische Ausdrücke;
7. chemische Formeln;
8. drucktechnisch hervorgehobene Wörter (Kursive, Majuskeln).

Weitere Differenzierungen bei der Kennzeichnung sind geplant, z.B. Eigennamen, geographische Namen u.ä..

Einer speziellen Konvention unterliegen die Satzzeichen:
Sie stehen immer zwischen Leerstellen, sind also innerhalb des Textkontinuums als Wort zu werten. Dadurch wird u.a. die eindeutige Zuordnung von Abkürzungs- und Satzschlußpunkt ermöglicht.

Die Zusatzcodierungen werden bewußt sparsam verwendet und auf solche Phänomene beschränkt, über deren Klassifizierung allgemein Einmütigkeit herrscht. Sie sind völlig wertfrei im Sinne einer zugrundeliegenden linguistischen Theorie - es werden z.B. keinerlei grammatische Strukturen

vermerkt, d.h. die aufgenommenen Texte sollen keinerlei Präjudizierung enthalten.

Eine gewisse Sonderstellung nimmt hier die Kennzeichnung der Nomina ein (bei Satzanfängen, die nicht durch ein Nomen besetzt sind, wird die Großschreibung ignoriert). Hier handelt es sich zugegebenermaßen um ein abfragbares Merkmal, das sich bei syntaktischen Untersuchungen als sehr nützlicher Indikator erweist.

Alle erwähnten Sonderkennzeichnungen können wahlweise bei dem Maschinenausdruck ausgegeben oder unterdrückt werden.

Übersicht über die zu beachtenden Schreibanweisungen für die Übertragung von Texten auf Lochstreifen - Stand vom 1.8.1968:

1. Textkopf

In die erste Zeile wird ttttttt und, ohne Abstand daran anschließend, das Symbol (d.h. die aus 3 Buchstaben bestehende Abkürzung für das entsprechende Wort) geschrieben. Dann folgt, mit einer Zeile Abstand, uuuuuu und ohne Leerstelle dahinter Verfasser, Titel und Jahresangabe. Bei Zeitungen wird jede Nummer mit ttttttt eingeleitet; dann folgt ohne Leerstelle das Symbol (3 Buchstaben) und als dreistellige Zeitungsnummer die jeweilige Nummer des Jahrestags. Nach einer Leerzeile schreibt man dddddd und direkt anschließend das volle Datum und die sonstigen Angaben im Kopf der Zeitung.

Beispiel für einen Buchtext:

tttttttLSO

uuuuuuErwin Strittmatter : Ole Bienkopp, 1963

Beispiel für einen Zeitungstext:

tttttZBZ061
ddddddMittwoch, 1. März 1967 . 16. Jahr. Nr. 51 . Druck in Hamburg .

2. Punktsetzung

Auch der Buchtitel ist ein Satz, muß also am Ende einen Punkt haben. Dasselbe gilt für alle Überschriften und Untertitel. Es gibt keine sprachlichen Äußerungen ohne Punkt.

3. Seitenangabe

Wo eine neue Seite des Quellentextes beginnt, schreibt man eine Leerzeile, dann ssssss, und ohne Zwischenraum die Seitenzahl (6stellig, rechtsbündig). Dann folgt eine weitere Leerzeile. Wird eine neue Zeitungsnummer begonnen, so steht nach tttttt... und dddddd... die Seitenangabe ssssss und die entsprechende Ziffer (6stellig, rechtsbündig). Bei jedem Übergang auf eine andere Seite muß ssssss..... geschrieben werden.

4. Zeileneinteilung

Die Zeilen der gedruckten Vorlage (mit Ausnahme von Zeitungstexten) werden beim Schreiben beibehalten. Wagenrücklauf bedeutet also: neue Zeile. Dabei müssen allerdings getrennte Wörter noch auf die vorhergehende Zeile geschrieben werden. Maximal gehen 100 Stellen (nicht Zeichen!) auf eine Zeile.

5 a. Leerstellen und Leerzeilen

Neue Absätze werden durch 6 Leerstellen eingeleitet; dann folgt ohne Leerstelle der Text. Bei neuen Abschnitten (Abstand im Originaltext, es fehlen eine oder mehrere Zeilen) wird eine Leerzeile (= 2mal Wagenrücklauf) geschrieben, dann wird wieder mit 6 Leerstellen eingeleitet; ebenso folgt am Ende eines Kapitels oder Artikels eine Leerzeile (vgl. Anweisg. 9).

5 b. Zeitungsartikel

Jeder Zeitungsartikel wird mit aaaaaa und der laufenden, von Hand in die Zeitung eingetragenen Nummer des Artikels (6stellig, rechtsbündig) eingeleitet.

6 a. Überschriften

werden mit uuuuuu eingeleitet. (Ohne Leerstelle weiter). Bei mehrzeiligen Überschriften muß am Anfang jeder Zeile uuuuuu stehen.

b. Beischriften

zu Bildern, Tabellen, Zeichnungen usw. werden mit b+ eingeleitet und mit +b abgeschlossen. Leerstellen wie bei Satzzeichen.

c. Quellenangaben,

z.B. "Bonn, den 28.1.1964 (dpa)" werden mit q+ eingeleitet und mit +q abgeschlossen. Leerstellen wie bei Satzzeichen.

d. Verfasserangaben

in Zeitungsartikeln (ausgeschrieben, abgekürzt oder als Siglen, z.B. "-k-") werden mit v+ eingeleitet und mit +v abgeschlossen. Leerstellen wie bei Satzzeichen.

e. Fremdsprachige Texte

(nicht Fremdwörter!) und eingeschobene Dialekttexte werden mit f+ eingeleitet und mit +f abgeschlossen. Leerstellen wie bei Satzzeichen.

f. Wörtliche Zitate

anderer Autoren werden mit z+ eingeleitet und mit +z abgeschlossen. Leerstellen wie bei Satzzeichen. Treffen mehrere Merkmale zusammen, so gilt die angegebene Reihenfolge.

7. Worttrennung

Am Ende einer Seite darf kein Teil eines Wortes stehen. Ist ein Wort in der gedruckten Vorlage über den Seitenwechsel hin abgesetzt, so wird es ganz auf die vorhergehende Seite genommen.

8. Satztrennung

Geht ein Satz über das Ende einer Seite weg, so ist nach dem Zeichen für die neue Seite (z.B. ssssss000321) in die nächste Zeile der Rest des Satzes zu schreiben.

9. Leerstellen

Vor und hinter jedem Wort, jeder Zahl (außer der laufenden Nummer) und jedem Satzzeichen ist eine Stelle frei zu lassen (aber nicht mehr als eine!).

10. Punkte

gelten nur als Satzzeichen, wenn sie einen Satz abschließen (man schreibt also 1. , 2. , 3. , 4. , 5. usw.). 3 Punkte (geschrieben : ...) sind ein Wort.

11. Apostrophe, Bindestriche, Abkürzungspunkte und Schrägstriche

sind keine Satzzeichen, haben also keine vorangehende Leerstelle. Sind durch Bindestrich verbundene Wörter durch "und", "oder" o.ä. unterbrochen (z.B. Wald- und Wiesen-Tee), so gehört der erste Bindestrich zum vorhergehenden Wort (WALD- UND WIESENTEE).
Bei Unterbrechung durch Anführungszeichen ("Entwicklungs"-Völker) wird der Bindestrich zweimal gesetzt ("ENTWICKLUNGS-" -VÖLKER).

12. Hervorhebung

Alle drucktechnisch hervorgehobenen Wörter (Kursiv, Majuskeln) sind am Ende mit einem Doppelpunkt (ohne Leerstelle) zu versehen.

13. Unterscheidung Buchstaben - Ziffern

Nie l (klein L) statt 1 (Eins), nie O (Buchstabe) statt 0 (Null) schreiben!

14. Groß- und Kleinschreibung. Außer den besonders zu kennzeichnenden Wörtern (vgl. u.a. Punkt 12) werden nur Nomina und Namen groß geschrieben. Dies gilt auch für Satzanfänge.

15. Potenzen
werden 10E5 (10^5), 2E3 (2^3) usw. geschrieben.

16. Sind Fragezeichen oder Ausrufezeichen satzschließend (folgt also ein großgeschriebenes Wort), so ist dahinter (zwischen Leerstellen) noch ein Punkt zu setzen. Diese Regel gilt auch für Gedankenstriche und für 3 Punkte ... .

17. Anführungszeichen:
Die Zeichen . " oder .» , ebenso .) usw. müssen vertauscht werden: " . oder» . , bzw.) .
Bei satzschließenden Ausrufe- oder Fragezeichen stehen die Anführungsstriche zwischen diesem Zeichen und dem Punkt (Originaltext: ! " oder ? " wird geschrieben: ! " . bzw. ? " .

Schreibtechnische Anweisungen

18. Am Anfang und Ende jedes Steifens muß etwa je ein Meter ungelochter Streifen (nur mit Führungslöchern) sein.

19. Abgerissene Streifen dürfen nicht mit Klebstoff oder Tesafilm, sondern nur mit den eigens dafür vorgesehenen Klebestücken geflickt werden.

20. Beim Umschalten auf Großbuchstaben darf der nächste (große) Buchstabe nicht zu schnell danach angeschlagen werden.

21. Der Wagenrücklauf darf nie von Hand betätigt werden, er erscheint sonst nicht auf dem Streifen. Vorsicht beim Neueinschalten der Maschine!

22. Insgesamt muß staccato geschrieben werden, d.h. die einzelnen Anschläge müssen deutlich zeitlich getrennt sein. Gleichmäßiges Tempo einhalten. Zu rasche Aufeinanderfolge von Anschlägen ergibt häufig Fehllochungen.

23. Folgende Abweichungen des IBM-Maschinenausdrucks vom Code des Streifenlochers sind zu beachten :

Streifenlocher		IBM-Maschine
!	=	/.
?	=	$
;	=	.,
: nach Leerstelle	=	..
: ohne Leerstelle am Wortende	=	+
"	=	,,
%	=	0/
ä	=	AE
ö	=	OE
ü	=	UE
ß	=	SS
§	=	blank (Leerstelle)

3. Textauswertung,

Der in der beschriebenen Weise aufbereitete Text ist in allen formalen Einzelheiten identifizierbar und kann nun nach beliebigen Gesichtspunkten im Rechner "verzettelt" werden. Dies geschieht im wesentlichen mit Hilfe des von G. Stickel entwickelten Programms INDEX (vgl. PI-11, DRZ, 64).

Das Bezugsmaß bildet jeweils eine Buchseite, deren Textzeilen fortlaufend gezählt werden (s.Abb. 1, Seite 24).
Aus technischen Gründen mußten anstelle der Maschinenausdrucke in den Abbildungen 1-6 Schreibmaschinenumschriften wiedergegeben werden, die dem Original im Schriftbild weitgehend angenähert sind.

3.1. Zerlegungen.

Zunächst werden einige Beispiele für solche Zerlegungen gegeben, bei denen das gesamte Textinventar vollständig, ohne Rest, aber jeweils unter wählbaren Aspekten wieder ausgegeben wird.

3.1.1. Wortformenregister.

Wortformenregister verzeichnen alle im Text vorkommenden Wortformen in alphabetischer Reihenfolge und enthalten Angaben zur Gesamthäufigkeit und Häufigkeit pro Seite. Es steht frei, bestimmte Typen von Wortformen (vgl. 2.3.1.) wie Satzzeichen, Formeln u.ä. von der Segmentierung auszuschließen (s. Abb. 2, Seite 25).

3.1.2. Rückläufige Register.

Die gleichen Register wie unter 3.1.1. vorgestellt können rückläufig sortiert gute Hilfe bei bestimmten Fragestellungen leisten (s. Abb. 3, Seite 26).

3.1.3. Häufigkeitsregister.

Das Häufigkeitsregister gibt das Gesamtvorkommen einer Wortform an sowie den Prozentanteil am Gesamtwortschatz des betreffenden Textes. Innerhalb einer Häufigkeitsgruppe werden die Formen alphabetisch sortiert (s. Abb.4, Seite 27).

3.1.4. Grundformenregister.

Als ein weiteres Hilfsmittel für verschiedenartige linguistische Untersuchungen erweist sich ein Register, in dem neben den vorkommenden Wortformen die zugehörigen (unflektierten) Grundformen aufgeführt sind. Hier können auch gewisse zusätzliche Informationen beigegeben werden. Diese Grundformenregister können im Gegensatz zu den bisher beschriebenen Registern in rein automatischem Verfahren erstellt werden.
Ein erstes solches Grundformenregister (für W.Heisenberg, Das Naturbild der heutigen Physik) wurde von D.Krallmann vom Institut für Phonetik und Kommunikationsforschung der Universität Bonn für das Institut für deutsche Sprache erarbeitet; weitere Register in teilweise abgewandelter Form werden folgen.

3.1.5. Zerlegung des Textes in Sätze.

Durch die Unterscheidung des Satzschlußpunktes von einem Abkürzungspunkt ist es möglich, den zwischen Leerstellen stehenden Punkt als Trennzeichen für die Segmentierung eines Textes in größere Einheiten als Wörter auszunutzen. Alles, was zwischen zwei solchen Punkten stehend vorgefunden wird, wird als Satz ausgegeben. Alle segmentierten Sätze der definierten Art werden mit fortlaufender Zählung und Bezug auf die Seite ihres Vorkommens im Buch aufgeführt (s.Abb.5, Seite 28).

3.2. Exzerpte.

Für Exzerpte bestimmter Teilstücke oder -mengen bietet sich einerseits die Möglichkeit, den Text nach bestimmten Elementen absuchen zu lassen, und zum andern das Verfahren, aus der Gesamtmenge der numerierten Sätze nur ganz bestimmte auszuwählen bzw. zu unterdrücken.

3.2.1. Kontextregister.

Bei den sogenannten Kontextregistern sind zwei Arten zu unterscheiden: a) aufgelistete Belegstellen für Wörter, b) aufgelistete Belegstellen für Endungen.

zu a) Es können Wörter beliebiger Anzahl gesucht werden. Jedes Wort wird mit Stellenangabe (Seite, Zeile) und Kontext, dessen Umfang beliebig wählbar ist, ausgegeben. Wählbar heißt: feste Anzahl von Wörtern vor und nach den Suchwort oder Suchwort plus Menge der Wörter, die zwischen anzugebenden Zeichen vorgefunden werden, z.B. ein Satz. Ausgegeben wird nach der alphabetischen Reihenfolge der Suchwörter; bei mehreren Fundstellen zu einem Wort nach der Reihenfolge ihres Vorkommens im Text.

zu b) Die Zahl der gesuchten Endungen ist nicht beschränkt; Stellenangabe und Kontextausgabe wie unter a). Die Sortierung erfolgt nach alphabetischer Ordnung der Endungen. Mehrere Funde zu einer gleichen Endung werden in alphabetischer Reihenfolge derjenigen Wörter, zu denen sie gehören, ausgegeben(s. Abb. 6, Seite 29).

Abb. 1: W. Heisenberg, Das Naturbild der heutigen Physik

MIT DEN GRUNDVORSTELLUNGEN* DER GRIECHISCHEN PHILOSOPHIE*
HERUMZUDENKEN . IN DIESER LAGE* HAT MIR DIE AUSBILDUNG* IM
PRINZIPIELLEN DENKEN* , DIE WIR AUF DER SCHULE* ERHALTEN HATTEN , AUSSERORDENTLICH
VIEL GEHOLFEN , MICH JEDENFALLS VERANLASST , NICHT MIT HALBEN
SCHEINLOESUNGEN* ZUFRIEDEN ZU SEIN , UND AUCH EINE GEWISSE KENNTNIS*
DER GRIECHISCHEN NATURPHILOSOPHIE* , DIE ICH MIR DAMALS ANGEEIGNET
HATTE , WAR MIR VON GROSSEM NUTZEN* .

WENN MAN IN DER HEUTIGEN ZEIT* UEBER DEN WERT* DER HUMANISTISCHEN
BILDUNG* SPRICHT , SO KANN MAN WOHL AUCH KAUM MEHR EINWENDEN ,
DASS DIE BEZIEHUNG* ZUR NATURPHILOSOPHIE* IN DER MODERNEN
ATOMPHYSIK* EIN EINMALIGER FALL* SEI UND DASS MAN SONST IN NATURWISSENSCHAFT* ,
TECHNIK* ODER MEDIZIN* MIT SOLCHEN PRINZIPIELLEN FRAGEN*
KAUM IN BERUEHRUNG* KOMME . DAS WAERE SCHON DESHALB FALSCH ,
WEIL VIELE NATURWISSENSCHAFTLICHE DISZIPLINEN* IN IHREN GRUNDLAGEN*
MIT DER ATOMPHYSIK* ENG VERBUNDEN SIND , ALSO SCHLIESSLICH AUF AEHNLICHE
GRUNDSAETZLICHE FRAGEN* FUEHREN WIE DIE ATOMPHYSIK* SELBST . DAS
GEBAEUDE* DER CHEMIE* ERHEBT SICH AUF DEM FUNDAMENT* DER ATOMPHYSIK* ,
DIE MODERNE ASTRONOMIE* HAENGT MIT IHR AUFS ENGSTE ZUSAMMEN
UND KANN OHNE ATOMPHYSIK* KAUM GEFOERDERT WERDEN , UND SELBST VON
DER BIOLOGIE* WERDEN SCHON BRUECKEN* ZUR ATOMPHYSIK* GESCHLAGEN . IN
DEN LETZTEN JAHRZEHNTEN* SIND IN VIEL HOEHEREM MASSE* ALS FRUEHER DIE
VERBINDUNGEN* ZWISCHEN DEN VERSCHIEDENEN NATURWISSENSCHAFTEN*
SICHTBAR GEWORDEN . AN VIELEN STELLEN* ERKENNT MAN DIE ZEICHEN* DES
GEMEINSAMEN URSPRUNGS* , UND DER GEMEINSAME URSPRUNG* IST SCHLIESSLICH
IRGENDWO DAS ANTIKE DENKEN* .

5. DER GLAUBE* AN UNSERE AUFGABE* .

MIT DIESER FESTSTELLUNG* BIN ICH NUN BEINAHE WIEDER ZUM AUSGANGSPUNKT*
ZURUECKGEKOMMEN . AM ANFANG* DER ABENDLAENDISCHEN KULTUR*
STEHT DIE ENGE VERBINDUNG* VON PRINZIPIELLER FRAGESTELLUNG* UND
PRAKTISCHEM HANDELN* , DIE VON DEN GRIECHEN* GELEISTET WORDEN IST .

Abb. 2: W. Bergengrün, Das Tempelchen

WORTREGISTER
==========

SEITE 1

NR.		GES.-VORK.	HAEUFIGK. PRO SEITE
1.	AB	3	000010. 1 000014. 1 000026. 1
2.	ABEND*	6	000024. 1 000025. 2 000026. 1 000032. 1 000042. 1
3.	ABENDS	1	000024. 1
4.	ABER	76	000008. 1 000009. 1 000010. 2 000011. 2 000012. 3 000013. 4 000014. 2 000015. 3 000016. 1 000017. 3 000018. 2 000019. 2 000020. 1 000021. 2 000022. 2 000023. 4 000025. 2 000026. 2 000027. 1 000028. 1 000029. 1 000030. 4 000031. 1 000032. 1 000033. 2 000035. 1 000036. 3 000037. 1 000039. 3 000040. 2 000041. 2 000042. 3 000043. 1 000044. 3 000045. 3 000046. 1 000047. 3
5.	ABGEBILDET	1	000038. 1
6.	ABGEBROCHEN	1	000009. 1
7.	ABGELEGT	1	000020. 1
8.	ABGELOEST	1	000034. 1
9.	ABGEMACHT	1	000023. 1
10.	ABGERISSEN	1	000014. 1
11.	ABGESUCHT	1	000039. 1
12.	ABGEWEHRT	1	000020. 1

Abb. 3: L.Mackensen, Deutsches Wörterbuch (zur Verfügung gestellt von Dr.Hübner, IBM)

M	/ MUNITIONSZUG	
M	/ FASSZUG	
M	/ EXPRESSZUG	
M	/ SANITAETSZUG	
M	/ RECHTSZUG	
M	/ GESICHTSZUG	
M	/ AUSZUG	
M	/ BODENAUSZUG	
M	/ KONTOAUSZUG	
M	/ KLAVIERAUSZUG	
M	/ KAISERAUSZUG	
M	/ STRAFREGISTERAUSZUG	
M	/ LUXUSZUG	
M	/ LASTKRAFTZUG	
M	/ SCHRIFTZUG	
M	/ DRAHTZUG	
M	/ NACHTZUG	
M	/ GELEITZUG	
M	/ PROSPEKTZUG	
M	/ LASTZUG	
M	/ FERNLASTZUG	
M	/ FESTZUG	
M	/ POSTZUG	
M	/ LAZARETTZUG	
M	/ BAUZUG	
M	/ REUZUG	
M	/ ZUZUG	
M	/ BLITZZUG	
M	/ KREUZZUG	
S	/ DEHUNGS-H	
S	/ AH	
	BAH	
	BAH	
W	/ HAGGADAH	
EW	GAH	
M	/ SCHAH	
M	/ PADISCHAH	
	IAH	
EW	JAH	
M	/ RAJAH	
S	/ DSCHAHELIJAH	
	ALLELUJAH	
W	/ BABLACH	
ZA	/ ALLAH	
	/ MASCHALLAH	
	INSCHALLAH	
M	/ MULLAH	
	MAEH	
EW	NAH	
UW	BEINAH	
EW	ERDENNAH	
EW	GEGENWARTSNAH	
EW	GOTTNAH	
EW	HERZNAH	
MV	/ ELOAH	
	PAH	
W	/ HAHNENKRAH	
W	/ HAHNENKRAEH	
W	/ OMRAH	
	/ KORAH	
S	/ JARRAH	
W	/ SIRRAH	
M	/ SURAH	
S	/ PASSAH	
S	/ LATAH	
EW	ZAEH	
S	/ GEZAEH	
S	/ BERGGEZAEH	
S	/ ACH	
	ACH	
M	/ BACH	
M	/ REBBACH	
M	/ WILDBACH	
M	/ MUEHLBACH	
M	/ ERLENBACH	
M	/ GLETSCHERBACH	
M	/ KREBSBACH	
M	/ GIESSBACH	
M	/ STAUBACH	
M	/ STURZBACH	
S	/ DACH	
S	/ ABDACH	
S	/ OBDACH	
S	/ LAUBDACH	

Abb. 4: M.Frisch, Homo Faber

HAEUFIGKEITSREGISTER

	HAEUFIGK.	PROZ.		HAEUFIGK.	PROZ.
ICH	2584	4.558	WIEDER	169	0.298
DIE	1181	2.083	AUS	155	0.273
SIE	1061	1.871	OHNE	154	0.272
UND	1032	1.820	ALLES	152	0.268
NICHT	1010	1.781	UEBER	148	0.261
ES	823	1.452	DU	146	0.258
DER	784	1.383	KEINE	144	0.254
ZU	783	1.381	MEIN	144	0.254
IN	723	1.275	VOR	140	0.247
DAS	603	1.064	HAT	137	0.242
WIE	602	1.062	WEISS	133	0.235
WAR	533	0.940	WEIL	125	0.220
MICH	485	0.855	EINMAL	123	0.217
EIN	454	0.801	IMMER	122	0.215
AUF	437	0.771	MEHR	119	0.210
IST	415	0.732	ZUM	118	0.208
VON	410	0.723	FUER	117	0.206
DASS	407	0.718	EINEN	116	0.205
MIT	405	0.714	EINEM	114	0.201
ALS	393	0.693	SCHON	111	0.196
ER	389	0.686	UNS	109	0.192
HANNA*	383	0.676	JA	106	0.187
DEN	380	0.670	ODER	104	0.183
WIR	368	0.649	AM	103	0.182
ABER	343	0.605	KEIN	103	0.182
EINE	337	0.594	IVY*	100	0.176
WAS	325	0.573	EINER	99	0.175
IM	311	0.549	SAGT	99	0.175
HATTE	303	0.534	BIN	97	0.171
IHR	302	0.533	NIE	97	0.171
MAN	301	0.531	IHN	93	0.164
MIR	291	0.513	HERBERT*	91	0.161
UM	291	0.513	MEINEN	91	0.161
SICH	274	0.483	WOLLTE	90	0.159
IHRE	264	0.466	WO	89	0.157
AN	246	0.434	DABEI	86	0.152
SO	235	0.414	KONNTE	85	0.150
DANN	233	0.411	SIND	85	0.150
NOCH	233	0.411	WUSSTE	85	0.150
MEINE	232	0.409	IHREN	84	0.148
NUR	230	0.406	MUSSTE	83	0.146
SAGTE	228	0.402	GING	82	0.145
DEM	225	0.397	MAEDCHEN*	82	0.145

28 000010 DIE UNIFORMEN* KANNTE ER NUR AUS FILMEN* .

29 000011 ETWAS MUERRISCH IM NACHMITTAEGLICHEN STAUB* UND GRASDUFT* FUHR ER WEITER AUF DER OSTDEUTSCHEN SEITE* DER AUTOBAHN* UND GRUEBELTE AN DER BEDEUTUNG* IHRER EINLADUNG* .

30 000011 ER HATTE SIE SEIT MEHREREN JAHREN* NICHT GESEHEN .

31 000011 SIE SCHICKTE IHM PROGRAMMHEFTE* UND FOTOGRAFIEN* ., ER VERGASS NICHT IHR SEINE BUECHER* ZU SCHICKEN .

32 000011 ERST IN DER LETZTEN ZEIT* HATTE SIE SICH OFFENBAR DARAN GEWOEHNT DASS ER IN SEINER ENTFERNUNG* VON SECHSHUNDERT KILOMETERN* GEDULDIG BEREIT WAR ZU AUSKUENFTEN* UEBER SEINEN TAGESLAUF* UND ZU GESPRAECHEN* UEBER DIE FREUNDE* , DIE SIE GEMEINSAM HATTEN AUS DER ZEIT* EINES MOEBLIERTEN ZIMMERS* IN EINER PARKSTRASSE* VON WESTBERLIN* .. ALS WOHNTEN SIE IN EINER STADT* NEBENEINANDER UND HAETTEN GLEICHE WORTE* FUER VERGLEICHBARES* .

33 000011 IHRE EINLADUNG* WAR BEILAEUFIG GEWESEN UND OHNE FREUNDLICHKEIT* UND ERKLAERT MIT NICHTS .

34 000011 ER HIELT IN DER SCHWEREN DAEMMERUNG* ZWISCHEN FREMDEN AUTOS* WIE ALLTAEGLICH UND STIEG AUS .

35 000011 DIE GEHSTEIGE* WAREN GERAEUMIG , KLEINKOEPFIGE PFLASTERSTEINE* IN REGENDUNKLEN FUGEN* , GROSSE ALTE BAEUME* MIT HALBOFFENEN KNOSPEN* .

36 000011 DER GEWICHTIGE RAUCHSCHWARZE STUCK* DER HAUSFRONTEN* HAETTE HELLER AUSGESEHEN , WAERE ER FRUEHER GEKOMMEN .

37 000011 BEKANNT WAR NOCH DAS KURZE SCHNAPPEN* DER WAGENTUER* , DANN KAM DAS HOHE TREPPENHAUS* GANZ AUS MARMOR* UEBER REINLICH ZERSCHLISSENEN TEPPICHBAHNEN* .

38 000011 SIE DRUECKTE DAS FENSTER* AUF UND SAH IHM BEIM AUSSTEIGEN* ZU ., VON OBEN HATTE SEIN WAGEN* EIN LANGES HERRSCHAFTLICHES AUSSEHEN* , UND ALS ER VOR DER TUER* GEBUECKT SIE ABSCHLOSS , SCHIEN ER ABSCHIED* ZU NEHMEN .

39 000011 ENTTAEUSCHT BEMERKTE SIE DAS MISSTRAUEN* , DAS IHN NACH WENIGEN SCHRITTEN* INNEHALTEN LIESS UND IN DER TASCHE* NACH DEN PAPIEREN* FUEHLEN , DIE SEINE ANWESENHEIT* ERLAUBTEN .

40 000011 ER HATTE SICH ABER NICHT UMGESEHEN , TRAT RASCH UND GLEICHMAESSIG AUF DIE HAUSTUER* ZU .

Abb. 6: M. Frisch, Homo Faber

FRAU KONTEXTREGISTER SEITE 1

000015. 1 ES WAR EINE F R A U E N S T I M M E * , ICH SCHWITZTE WIEDER UND MUSSTE MICH SETZEN , DAMIT MIR NICHT SCHWINDLIG WURDE , MAN KONNTE MEINE FUESSE* SEHEN .

000035.14 MAN MUSSTE FAST SCHREIBEN , BLOSS DAMIT DIE LIEBEN LEUTE* NICHT FRAGTEN , OB MAN DENN KEINE F R A U * HABE , KEINE MUTTER* , KEINE KINDER* , - ICH HOLTE MEINE HERMES-BABY* (SIE IST HEUTE NOCH VOLL SAND*) UND SPANNTE EINEN BOGEN* EIN , BOGEN* MIT DURCHSCHLAG* , DA ICH ANNAHM , ICH WUERDE AN WILLIAMS* SCHREIBEN , TIPPTE DAS DATUM* UND SCHOB - PLATZ* FUER ANREDE* .. " MY DEAR* /.

000036. 2 ICH KONNTE SIE NICHT EINMAL UM ZUSTELLUNG* VON FILMEN* BITTEN UND WAR MIR BEWUSST , DASS IVY* , WIE JEDE F R A U * , EIGENTLICH NUR WISSEN MOECHTE , WAS ICH FUEHLE , BEZIEHUNGSWEISE DENKE , WENN ICH SCHON NICHTS FUEHLE , UND DAS WUSSTE ICH ZWAR GENAU .. ICH HABE HANNA* NICHT GEHEIRATET , DIE ICH LIEBTE , UND WIESO SOLL ICH IVY* HEIRATEN $

000037. 9 ABER DASS ICH DARAN DACHTE , IHREN STUDEBAKER* ZU VERKAUFEN , DAS FAND SIE UNMOEGLICH , BEZIEHUNGSWEISE TYPISCH FUER MICH , DASS ICH NICHT EINE SEKUNDE* LANG AN IHRE GARDEROBE* DAECHTE , DIE MIT DEM HIMBEER-STUDEBAKER* STAND UND FIEL , TYPISCH FUER MICH , DENN ICH SEI EIN EGOIST* , EIN ROHLING* , EIN BARBAR* IN BEZUG AUF GESCHMACK* , EIN UNMENSCH* IN BEZUG AUF DIE F R A U * .

000038. 3 " HANNA - SEINE F R A U * " .

000071. 1 (SICHER WAR ICH BEI F R A U E N* NIE .

000077. 3 " SAGTE SIE - NICHT NUR VERSTAENDNISLOS , WIE ICH'S VON F R A U E N * GEWOHNT BIN , SONDERN GERADEZU SPOETTISCH , WAS MICH NICHT HINDERTE , DAS APPARAETCHEN* VOLLKOMMEN ZU ZERLEGEN ., ICH WOLLTE WISSEN , WAS LOS IST .

000080.24 EINER STREIKTE , ALS ER HOERTE , DASS EINE F R A U * ZUGEGEN WAERE ., DAS WAR IHM ZUVIEL ODER ZUWENIG .

000099.15 SABETH* WAR SCHON EINE RICHTIGE F R A U * , WENN SIE SO LAG , KEIN KIND* ., ICH NAHM EINE DECKE* VOM OBEREN BETT* , DA SIE VIELLEICHT FROR , UND DECKTE SIE ZU .

3.2.2. Satzlisten.

Für bestimmte Untersuchungsgegenstände, die nicht "maschinen-explizit" sind (vgl. Punkt 4), für die ein automatischer Suchlauf also nicht durchführbar ist, läßt sich dennoch über die abzurufenden Satznummern eine sinnvolle Auswahl aus einer Textmenge treffen. Jeder Text kann so in jede beliebige Menge ganz spezieller Arbeitslisten verzettelt werden, in denen jeder Satz mit seiner Stellenangabe (Nummer und Seite im Text) versehen ist.

4. Das Verfahren "Parallelcodierung" (von Alex Ströbl).

4.1. Begründung des Verfahrens.

Wenn im Institut Texte über Lochstreifen auf Magnetband übertragen werden, dann aus zwei Gründen:

a) Die Texte interessieren nicht um ihrer selbst willen, sondern sie sind vielmehr ausgewählte Repräsentanten aus der Klasse aller deutsch geschriebenen Texte der Gegenwart. Sie stellen eine Menge von Trägern bestimmter Merkmale dar, nämlich von Merkmalen, die für die Untersuchung der deutschen Gegenwartssprache als relevant angesehen werden. Es geht nicht um z.B. "Solange dein Herz schlägt" von Pia Stauffen, sondern um die Subjekte, Konjunktive, Satzbaupläne, Wortstellungserscheinungen etc., die sich in diesem "Schicksalsroman" finden; denn was Pia Stauffen schreibt, gilt als Gegenwartsdeutsch und kann somit als Grundlage einer entsprechenden Untersuchung dienen.

b) Die Texte sollen maschinell bearbeitbar sein, damit die Möglichkeiten der Arbeitsersparnis, die moderne Datenverarbeitungsanlagen bieten, ausgenützt werden können. So ist es unter Umständen rationeller, sich z.B. seine Exzerpte über den Computer machen zu lassen, als selbst Karteikarten zu schreiben.
Diese maschinelle Bearbeitbarkeit findet nun dort ihre Grenze, wo die interessierenden Merkmale bzw. ihre Träger vom Computer nicht identifiziert werden können, d.h. wo das Merkmal so beschaffen ist, daß es (zur Zeit) kein Programm gibt, das den Träger dieses Merkmals feststellt. Das Merkmal ist zwar für den untersuchenden

Linguisten "explizit", für die Maschine aber nur "implizit" in dem Text enthalten. Eingesetzt werden kann jedoch die Maschine nur bei der Bearbeitung von Merkmalen, die "maschinen-explizit" sind.
In welchem Ausmaß die auf Band befindlichen Texte des Instituts für die laufende Institutsarbeit von Nutzen sein können, hängt also davon ab, wieweit es möglich ist, Merkmale, die bei diesen Arbeiten interessieren, maschinenexplizit zu machen.

Maschinen-explizit sind zur Zeit nur diejenigen Merkmale, die sich als Folgen von n alphamerischen Zeichen ($n \geq 1$) oder als Kombinationen solcher Folgen darstellen lassen, d.h. Fälle, in denen Eigenheiten der graphematischen Darstellung des Merkmalträgers als Merkmal gesucht werden. Darauf beruhen Wortsuchprogramme, Ausgabe bestimmter Sätze aufgrund von Satznummern und Ähnliches. Nicht faßbar sind alle diejenigen Merkmale, die ein "Verstehen" des Textes voraussetzen. Hierher gehören Merkmale wie "Subjekt", "Dativ Plural", "Satzglied".

Zwei Wege gibt es, die da weiterführen könnten:

a) Entwicklung von Algorithmen, die zum gleichen Ergebnis führen wie das "Verstehen". Solche sog. "Analyseprogramme" setzen jedoch schon die Ergebnisse der Untersuchung der betreffenden Erscheinung voraus, denn erst auf der Grundlage ziemlich ausgedehnten Wissens über die Erscheinung können sie geschrieben werden. Dieser Weg bietet also keine Hilfe für laufende Untersuchungen.

b) Zusätzliche Kennzeichnung der Merkmalsträger in maschinen-expliziter Form. Das erstmalige Feststellen der Merkmalsträger wird im Rahmen der laufenden Arbeit vorgenommen, bietet also noch keine Arbeitsersparnis; dann aber können die eingegebenen Kennzeichnungen als "sekundäre" explizite Merkmale betrachtet werden und die Maschine ist voll einsetzbar. (Das Verfahren ist ähnlich dem, daß ein Fachmann in einem Text die ihn interessierenden Stellen anstreicht, und ein Laie übernimmt das Herausschreiben der Exzerpte).

Ein Einfügen der sekundären Merkmale in den laufenden Text empfiehlt sich nicht, denn dann müßte bei jedem Hinzukommen eines Satzes sekundärer Merkmale und bei jeder Korrektur schon vorhandener Merkmale das Text-Band neu erstellt werden; außerdem würde dadurch eine Arbeit mit den Merkmalen unabhängig vom Text sehr erschwert.

4.2. Beschreibung des Verfahrens.

Aufgrund dieser grundsätzlichen Überlegungen wurde das im Folgenden zu beschreibende Verfahren ("Parallelcodierung") entwickelt. Die Arbeit mit diesem Verfahren steckt im Institut allerdings erst in den Anfängen, so daß noch nicht über Erfahrungen damit berichtet werden kann. Ein sehr ähnliches System wird jedoch seit Jahren bei Prof. Eggers in Saarbrücken mit gutem Erfolg praktiziert.[1] Bei der "Parallelcodierung" werden die sekundären Merkmale nicht dem Text eingefügt, sondern zugeordnet. Für die Zuordnung werden explizite primäre Merkmale des Textes ausgenützt: Die "Text-Sätze" (= was zwischen zwei Punkten steht) werden durchnumeriert ("Satzzerlegung"), und weiter können auch die "Text-Wörter" (= was zwischen zwei Leerstellen steht; nach unseren Schreibkonventionen auch die Satzzeichen) eines Satzes ebenfalls durchnumeriert gedacht werden. Dann ist jedes Text-Wort eindeutig bestimmt durch die Kennzeichnung des Textes, die Nummer des Text-Satzes, in dem es steht, und die Nummer, die es in diesem Satz hat: Text-Wort Homo Faber 23,3 wäre beispielsweise das 3. Text-Wort des 23. Text-Satzes von Frischs "Homo Faber". Durch die Angabe von Text, Text-Satz-Nummer und Text-Wort-Nummer ist also eine eindeutige Zuordnung eines sekundären Merkmals zu einem Text-Wort möglich.

Da die Grundlage der Zuordnung das einzelne Text-Wort ist, ist das Verfahren dann am rationellsten, wenn als Merkmalsträger einzelne Text-Wörter oder kleinere Gruppen von Text-Wörtern in Frage kommen ("Subjekt", "Dativ Plural"); mit dem Anwachsen der Anzahl von Text-Wörtern, die

zusammen Merkmalsträger sind, verliert es an praktischem Wert (etwa ganzer Satz Merkmalsträger : "Fragesatz").[2] Nicht anwendbar ist das Verfahren, wenn das Merkmal nicht so definiert werden kann, daß seine Träger Text-Wörter oder Gruppen von Text-Wörtern sind.

Die Menge der sekundären Merkmale und ihrer Zuordnungen wird als offen betrachtet : Jeder, der mit den auf Band befindlichen Texten arbeitet, erstellt diejenigen Daten, die für seine Arbeit nützlich sind. Bei jeder späteren Arbeit kann dann auf die Daten aus früheren Untersuchungen zurückgegriffen werden, soweit sie für die neue Fragestellung von Interesse sind. Wurde etwa einmal eine Segmentierung nach Satzgliedern durchgeführt, dann kann bei einer Untersuchung der Nominalgruppe teilweise darauf aufgebaut werden, denn Subjekt und Objekte sind häufig Nominalgruppen.

Um eine möglichst gute derartige Ausnützung des schon Erarbeiteten zu ermöglichen, wurden einige einheitliche Konventionen geschaffen :

a) Festgelegt ist die Einteilung der Lochkarte, die als Datenträger dient. So können Programme, die für die Auswertung eines bestimmten Datensatzes geschrieben wurden, - soweit das von der jeweiligen Fragestellung her möglich ist - immer wieder verwendet werden.

b) Festgelegt sind auch die Textstücke, die von jedem, der das Verfahren anwendet, bearbeitet werden sollen[3]. Dadurch wird dem vorgebeugt, daß die einzelnen Merkmale jeweils verschiedenen Textstücken zugeordnet werden, und erreicht, daß zu diesen Textstücken alle bisher erarbeiteten Daten zur Verfügung stehen. (Selbstverständlich steht es dem einzelnen frei, mehr Text als festgelegt zu codieren; es geht nur um das gemeinsame Minimum).

5. Übersicht über die Textaufbereitung - Stand vom 1.10.1968:

Werk	auf Lochstreifen	auf Magnetband	Zerlegungen **) W	H	S	R	Zugriff möglich *)	Umfang S = Buchseite
LBT	+	+	+	+	+			48 S.
LBC	+	+			+		+	303 S.
LFH	+	+	+	+	+			252 S.
LGB	+	+			+		+	493 S.
LMB	+	+	+	+	+		+	127 S.
LSO	+	+	+	+			+	365 S.
LJA	+	+			+		+	337 S.
TJM	+	+	+	+	+			62 S.
TPM	+	+			+		+	180 S.
TSH	+	+						62 S.
WBO	+	+			+		+	263 S.
WBM	+	+	+	+	+		+	238 S.
WGW	+	+			+		+	134 S.
WGS	+	+						237 S.
WHK	+	+	+	+	+		+	113 S.
WHN	+	+	+	+	+			46 S.
WJA	+	+						501 S.
WJZ	+	+						
WPE	+	+			+		+	449 S.
WSP	+	+			+		+	256 S.
WUB	+	+			+		+	160 S.
MHE	+	+			+		+	448 S.
ZFA	+	+			+		+	1 Monat 1966
ZWE	+	+						1 Monat 1965, 2 Monate 1966
ZBW	+	+	+	+	+		+	Heft 1,2 u.3, 1967
ZSG	+	+	+	+	+		+	Heft 12, 1966
ZUR	+	+			+		+	Heft 11, 1966; Heft 1, 1967
ZBZ	+	+					+	7 Monate 1967

*) Diese Werke bedürfen noch einiger Nachkorrekturen, sind aber für bestimmte Fragestellungen schon maschinell erschließbar.

**) W = Wortformenregister, H = Häufigkeitsregister, S = Satzzerlegung, R = Rückläufige Register.

6. Anhang.

6.1. Sonstige laufende Arbeiten.

Programme, die über die Textaufbereitung hinausgehen, dienen vorläufig hauptsächlich der statistischen Auswertung der Texte. So existieren spezielle Zählprogramme, numerische Berechnungen verschiedenster Art und - als häufigstes Arbeitsmittel benutzt - immer wieder Sortierprogramme.

6.2. Austauschbarkeit von Daten, Programmen usw.

Die Frage nach der Kompatibilität ist unbedingt zu spezifizieren. Wir unterscheiden :

a) Kompatibilität der Daten, wobei zu beachten ist, daß die Daten sowohl auf Lochstreifen als auch auf Magnetbändern gespeichert sind;

b) Kompatibilität der Programme, wobei zu berücksichtigen ist, daß diese sowohl problemorientiert als auch maschinenorientiert sind.

Eine völlig generelle und problemlose Austauschbarkeit gibt es in beiden Fällen nicht. Die Codierung der Lochstreifen, der Schreibmodus der Bänder, der Grad der Maschinengebundenheit von Programmen und Speicherung setzen hier Grenzen, lassen sich jedoch in vielen Fällen durch Konvertierungsprogramme adaptieren.

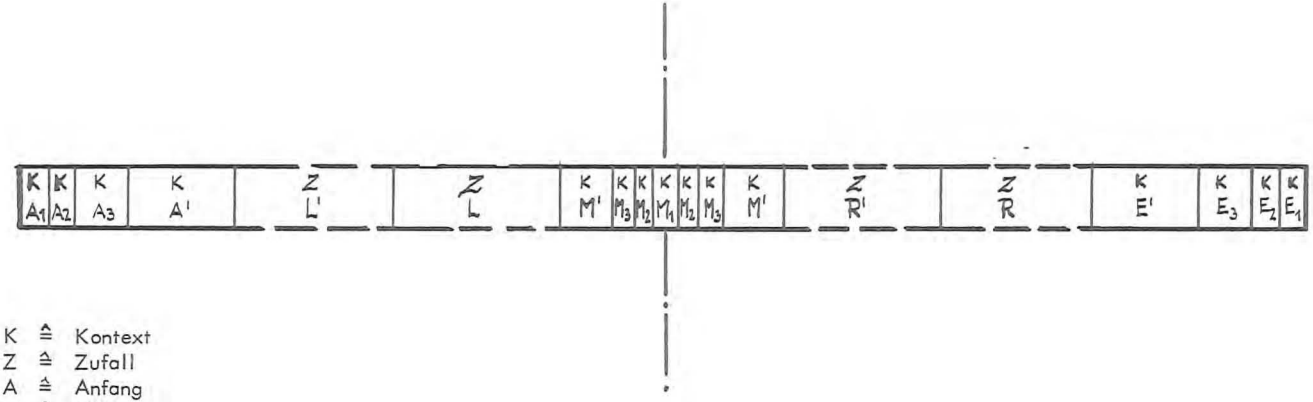
Anmerkungen zu Teil I

1) Vgl. die Arbeitsberichte über "Linguistische Arbeiten des Germanistischen Instituts und des Instituts für Angewandte Mathematik der Universität des Saarlandes", hg. v. Hans Eggers und Johannes Dörr, die hektographiert in zwangloser Folge erscheinen.

2) Hat man es im Zusammenhang mit einer bestimmten Fragestellung mit immer den gleichen Gruppen von Text-Wörtern als Merkmalsträgern zu tun, (beispielsweise bei einer Untersuchung der Satzglieder nach verschiedenen Gesichtspunkten), so empfiehlt sich die folgende Abwandlung des Verfahrens: In einem ersten Schritt werden diese Gruppen bestimmt und den einzelnen Text-Wörtern das Merkmal "Zugehörigkeit zu der Gruppe X" zugeordnet (man kann sich etwa die Gruppen im Satz durchnumeriert denken und jedem Text-Wort die Nummer der Gruppe, zu der es gehört, als sekundäres Merkmal zuordnen). Alle weitere Zuordnung von sekundären Merkmalen zu den konstanten Segmenten geschieht als Zuordnung nicht zu den Text-Wörtern, sondern zu den im ersten Schritt in ihrer Zusammengehörigkeit definierten Gruppen. Der Bezug zu den einzelnen Text-Wörtern bleibt möglich, denn die Zugehörigkeit der Text-Wörter zu den Segmenten ist ja gegeben.

3) Das Auswahlverfahren für die Textstücke wurde unter dem Gesichtspunkt bestimmt, daß diese Textstücke so beschaffen sein sollen, daß sie als Arbeitsgrundlage für möglichst viele und möglichst verschiedene Fragestellungen dienen können sollen. Wichtige Hilfe bei der Definition des Auswahlverfahrens ist Herrn stud. rer. pol. Werner Müller, Mannheim, zu verdanken. Von jedem Text wird gleich viel genommen und nach den gleichen Kriterien. Grundlage für alle Bestimmungen sind die Text-Wörter unter Nichtberücksichtigung der Satzzeichen und die Text-Sätze. Für die "Parallelcodierung" als erstes Korpus dienen aus jedem Text möglichst

genau 4200 Text-Wörter; grundsätzlich gilt, daß jeder angefangene Satz vollständig bearbeitet werden soll. Davon werden 2100 Text-Wörter in Gruppen von je einem Text-Satz aus zwei bestimmten Gebieten des Textes (s.u.) nicht fortlaufend zufällig bestimmt ("Zufall"), die anderen 2100 Text-Wörter in drei in sich geschlossenen Blöcken ("Kontext") zu je 700 von Anfang, Mitte und Ende genommen.
Das so bestimmte Material wird nocheinmal in drei Dringlichkeits-Stufen aufgeteilt, die genau wie das Gesamtmaterial zusammengesetzt sind, jedoch nur 1/4 (Stufe I), 1/4 (Stufe II) und 1/2 (Stufe III) davon ausmachen. Die Lage dieses Korpus im Gesamttext ist folgende:



K ≙ Kontext
Z ≙ Zufall
A ≙ Anfang
M ≙ Mitte
E ≙ Ende
L ≙ links
R ≙ rechts

Die Indexzahlen deuten an, in welcher Stufe ein bestimmtes Stück Text bearbeitet werden soll. _' deutet Textstücke an, welche nach dem Dreistufenplan unberührt bleiben, um für eventuelle spätere Ausdehnung der Erhebungsmengen zur Verfügung zu stehen.

Stufe I			Stufe II			Stufe III		
KA_1	≙	175 Wörter	KA_2	≙	175 Wörter	KA_3	≙	350 Wörter
KM_1	≙	175 Wörter	KM_2	≙	175 Wörter	KM_3	≙	350 Wörter
KE_1	≙	175 Wörter	KE_2	≙	175 Wörter	KE_3	≙	350 Wörter
aus ZL	≙	263 Wörter	aus ZL	≙	262 Wörter	aus ZL	≙	525 Wörter
aus ZR	≙	262 Wörter	aus ZR	≙	263 Wörter	aus ZR	≙	525 Wörter
		1050 Wörter			1050 Wörter			2100 Wörter

insges. 4200 Wörter

II. Zur Dokumentation und maschinellen Bearbeitung von Zeitungstexten in der Außenstelle Bonn

von Manfred W. Hellmann

Übersicht

Anhänge :

Allgemeiner Teil : Zur Dokumentation bei Zeitungstexten

1. Ausgangslage

Als die Außenstelle Bonn im August 1964 gegründet wurde, erhielt sie die Aufgabe zugewiesen, die seit der politischen Teilung Deutschlands eingetretenen Veränderungen der deutschen Sprache in den beiden Teilen Deutschlands zu beobachten, zu registrieren und zu analysieren. Es bestand Klarheit darüber, daß diese Aufgabe nur Teil der größeren, die deutsche Sprache der Gegenwart überhaupt wissenschaftlich zu erforschen, sein kann.

Als Materialgrundlage für alle Untersuchungen kamen von vorn herein nur gedruckte, veröffentlichte Texte in Betracht, da uns die Beschaffung von Material gesprochener Sprache aus der DDR, vor allem von spontan gesprochener Umgangssprache, jedenfalls in methodisch ausreichender Breite aus bekannten Gründen nicht möglich ist. [1)]

Die Untersuchungen sollten sich vor allem, wenn auch nicht ausschließlich, auf den Wortschatz richten, und zwar auf den Wortschatz allgemein, sofern er überhaupt für sprachliche Veränderungen in Betracht kommt. Es sollte jedenfalls nicht ein bestimmtes Sachgebiet von vorn herein bevorzugt werden.

Das Textmaterial sollte aktuell sein und weder zu sehr von einigen oder einem bestimmten Verfasser stammen noch ausgeprägt fachsprachlichen Charakter haben.

Schließlich sollte das Material so gewählt werden, daß es ost-west-vergleichende Studien zuläßt, d.h. es sollte eine nach Zweck und Ziel seiner Entstehung, Aufbau, Erscheinungs- und Verbreitungsart möglichst vergleichbare Struktur besitzen.[2)] Diesen Bedingungen zufolge entschied sich die Außenstelle für die Aufnahme von Zeitungstexten, und zwar von Texten aus Tageszeitungen großer Verbreitung und überregionaler Bedeutung.

2. Allgemeines zum Medium "Tageszeitung"

Tageszeitungen der bezeichneten Art weisen eine Reihe gemeinsamer charakteristischer Züge auf, die den vorher genannten Bedingungen entsprechen :

2.1. Heterogenität im einzelnen

1. Vielheit von Verfassern

An jeder Zeitungsausgabe arbeitet eine Vielzahl von Redaktionsmitgliedern, Korrespondenten und anderen Autoren mit; theoretisch sind ebenso viele individuelle Schreibgewohnheiten zu erwarten. Ausgleichend wirken dagegen: der Einfluß der Redaktionsleitung oder einer ihr übergeordneten Instanz, die normierende Kraft des "teamworks" ("Kollektivs") in der Redaktion selbst, der Einfluß der Agenturen (in der DDR ADN, in der BRD vor allem dpa, ap und upi).

2. Vielfalt von Sachgebieten

Theoretisch ist die Zahl der Sachgebiete, die in einer Zeitung Berücksichtigun finden können, nahezu unbegrenzt groß. Allerdings wird jede Zeitung bestimmte Sachgebiete mehr pflegen als andere; allgemeine Tageszeitungen also überwiegend Sachgebiete von allgemeinerem Interesse. Die Hauptsachgebiete, die in nahezu jeder Ausgabe jeder großen Tageszeitung vertreten sind, werden in "Sparten" oder Ressorts zusammengefaßt; die Gliederung und Bezeichnung dieser Ressorts (am häufigsten Politik, Wirtschaft, Sport, Kulturelles, Unterhaltung, Lokales o.ä.) ist wie ihr Umfang von Zeitung zu Zeitung verschieden Auch innerhalb einer Zeitung können die Sparten nach Rang und Breite der Behandlung wechseln [3]. Die Hauptsachgebiete fächern sich in eine systematis nur schwer zu fassende Fülle von engeren und engsten Sachgebieten auf [4]. Überschneidungen mehrerer Sachgebiete sind häufig. Sie treten besonders dann auf, wenn Zeitungen eine Sparte besonders pflegen: eine stark politisch orientierte Zeitung (wie das ND) behandelt auch kulturelle oder wirtschaftliche Zusammenhänge unter einem politischen Aspekt, bei Wirtschaftszeitungen verhält es sich entsprechend anders.

3. Vielfalt an Themen [5]

Bedingt durch das ihr wesensgemäße Streben nach Aktualität, durch ihre Bindu an den Strom der Zeit ist die Zeitung einem ständigen Wechsel der Themen unterworfen bzw. auf diesen angewiesen. Nur sehr wenige Themen können sich

über längere Zeit hin halten [6]. Die Zahl der Themen muß als beliebig groß angesetzt werden; sie wird begrenzt ausschließlich durch den unterschiedlichen Informations- (Propaganda-, Werbe-, Sensations- usw.) -wert, den die Redaktion den Ereignissen, Sachverhalten oder Gegenständen, die ihr zur Kenntnis kommen, zumißt.

4. Mehrzahl an Zielen oder Zwecken des Schreibens

In den verschiedenen "Artikeln"[7] einer Zeitung werden ersichtlich verschiedene, wechselnde Ziele oder Zwecke verfolgt. Neben den "klassischen" publizistischen Zielen der Unterrichtung (Information), Beeinflussung (Meinungs- und Verhaltensbildung), Belehrung und Unterhaltung, die allerdings im Einzelfall selten rein in Erscheinung treten, könnte man noch Werbung (Bedarfsweckung und Bedarfslenkung) und Aufruf (Aufforderung zum Handeln oder bestimmten Verhaltensweisen als publizistische Ziele eigenen Charakters definieren [8].

5. Viefalt an Schreib- bzw. Mitteilungsformen [9]

Schon der im engeren Sinne redaktionelle Teil einer Zeitung weist eine Fülle verschiedener "Artikelformen" auf (Leitartikel, Kommentare, Berichte, Meldungen, Glossen usw.), daneben finden sich zahlreiche Sonderformen, die teils von der Redaktion, teils von privaten Auftraggebern zu verantworten sind, wie Romanabdrucke, Leserbriefe, Anzeigen, Tabellen, Wetterberichte, Abdrucke von Reden, Dokumenten, amtlichen Bekanntmachungen usw. Sie alle sind außerdem nach Umfang und Aufmachung (Schriftgrad der Überschrift, Spaltenzahl, Plazierung) unterschieden.

Die hier angedeutete außerordentliche Heterogenität, die für alle großen Tageszeitungen kennzeichnend ist, macht es wahrscheinlich, daß das Material auch sprachlich vielfältig, vielschichtig differenziert ist.[10] Sie stellt den Bearbeiter allerdings auch vor besondere Probleme, wenn er trotzdem allgemeingültige Aussagen machen will.

2.2. Ambivalenz und Homogenität im ganzen

Besondere Eigentümlichkeiten des Mediums "Tageszeitung" ergeben sich aus ihrer Eigenschaft als Massenkommunikationsmittel; sie haben, ebenso wie die Vielfalt in thematischer, inhaltlicher und formaler Hinsicht, große Bedeutung auch für die Sprache (Sprachgebrauch) der Zeitung.

Zwei Gesichtspunkte sind zu unterscheiden:

2.2.1. Zeitungen sind O r g a n für die (einheitliche oder nicht einheitliche) Meinung oder die Absichten d e r R e d a k t i o n bzw. für die Meinung oder Absichten derer, denen die Redaktion die Spalten ihrer Zeitung öffnet, etwa des Herausgebers, der Regierung, politischer, wirtschaftlicher, sozialer, weltanschaulicher (Interessen-) Gruppen bzw. der d o m i n i e r e n d e n oder allein herrschenden unter diesen G r u p p e n (wie der SED-Führungsgruppe in der DDR.) Jede Zeitung ist also Verbreitungsmittel für die Seh-, Denk- und damit auch Redeweise der Redaktion bzw. der sie beeinflussenden Gruppen oder Gruppe. Sie wird auch sprachlich deren Eigentümlichkeiten widerspiegeln. Der Grad der Beeinflussung und damit der gruppensprachlichen Eigentümlichkeiten ist allerdings sehr verschieden; er wird dort gering sein, wo in einer Zeitung bzw. in der Redaktion mehrere Gruppenmeinungen konkurrieren; er wird da am größten sein, wo eine Gruppe nicht nur eine bestimmte Zeitung, sondern die veröffentlichte Meinung überhaupt beherrscht, wo die Zeitung sich selbst als Verbreitungsmittel dieser einen Meinung versteht.

2.2.2. Zeitungen müssen aber selbst dann, wenn sie sich im bezeichneten extremen Sinn als Verbreitungsmittel eines Meinungsmonopols verstehen, Rücksicht nehmen auf die Aufnahmebereitschaft und Aufnahmefähigkeit ihrer Leserschaft, d.h. bei weit verbreiteten Tageszeitungen: einer breit gestreuten und differenzierten Leserschaft [11]. Im Normalfall müssen sie allgemeinverständlich bleiben und dürfen nur bis zu einem gewissen Grade enger fach-

oder gruppensprachlichen Einflüssen nachgeben. Der Grad der Rücksichtnahme und des Bemühens um Allgemeinverständlichkeit ist auch hier verschieden; er wechselt nicht nur von Zeitung zu Zeitung, sondern auch innerhalb der Sparten und Ausgaben derselben Zeitung: man läßt gelegentlich bewußt auch enger fachlich usw. gebundene Redeweisen zu, um den Ansprüchen einer differenzierten Leserschaft gerecht zu werden, meist jedoch ohne ein einmal gewonnenes mittleres Niveau zu sehr zu überschreiten.

Jede große Tageszeitung enthält also - auch sprachlich - beide Komponenten. Die jeweilige Position zwischen den beiden extremen Möglichkeiten bestimmt jedoch wesentlich den individuellen Charakter einer Zeitung.

Diese doppelte Eigenheit der Zeitung, verbunden mit der ihr ebenso eigentümlichen Verpflichtung zur Aktualität, läßt die Zeitung als besonders geeignetes Medium für Untersuchungen im Rahmen des der Außenstelle gestellten Arbeitsthemas erscheinen. Je nachdem, ob im Rahmen spezieller Untersuchungen einmal mehr Gewicht auf die Komponente der politisch-geistig werbenden und führenden Sprache oder mehr auf die der allgemein kommunikativen Sprache gelegt wird, müssen oder können dann noch Texte weniger ambivalenten Charakters hinzugezogen werden.

3. Wissenschaftliche Fragestellungen

(Einzelanalysen und Dokumentation)

Das der Außenstelle gestellte Rahmenthema verlangt sowohl die Untersuchung einer ganzen Anzahl verschiedenster, teils engerer, teils weiterer Spezialgebiete, als auch Sammlungen und Untersuchungen zum Wortschatz allgemein. Für alle diese, im einzelnen noch nicht übersehbaren, Untersuchungen mußte das Material erst beschafft und bereitgestellt werden. So weit wie möglich sollten für diese Zwecke die Mittel der elektronischen Datenverarbeitung eingesetzt werden.

Es war allerdings klar, daß die Außenstelle weder den Ehrgeiz noch auch je die Möglichkeit haben würde, die im Rahmen des Gesamtthemas anfallenden einzelnen Aufgaben selbst zu bearbeiten. Der Zielsetzung des Instituts entsprechend sollte daher damit begonnen werden, nicht nur für den Bedarf der Außenstelle selbst und des Instituts, sondern auch anderer interessierter Forscher Material bereitzustellen. Gerade im Hinblick auf die Benutzung durch andere Wissenschaftler mußte die Möglichkeit für eine möglichst große Zahl verschiedener Fragestellungen geschaffen werden, und zwar derart, daß, sofern sich die Fragestellung aus dem aufgenommenen Material nicht voll beantworten läßt, wenigstens ein Grundmaterial geliefert werden kann, das dann vom betreffenden Forscher durch spezieller ausgewähltes Material ergänzt werden kann. Damit trat neben und vor die Aufgabe, Einzelanalysen zu erstellen, die Aufgabe der Dokumentation [12]. Entsprechend der vorher erläuterten Beschränkung auf das Medium "Zeitungen" handelt es sich vorerst um Dokumentation bei Zeitungstexten.
Die folgenden Ausführungen beziehen sich grundsätzlich auf diese.

4. Dokumentation bei Zeitungstexten

4.1. Allgemeine Bedingungen

An eine Dokumentation im bezeichneten Sinne werden bestimmte Anforderungen gestellt:

1. Die Texte müssen in ausreichender Vielfalt aufgenommen werden. Diese Bedingung wird durch die Eigenart des Mediums "Zeitung" erfüllt (siehe oben).
2. Die Texte müssen in ausreichend großer Menge aufgenommen werden, und zwar in so großer, daß
   a) Zufälligkeiten in der Zusammenstellung der einzelnen Textteile (Teiltexte) ausgeglichen werden,

b) auch zur Bearbeitung speziellerer Fragestellungen noch genügend Material zur Verfügung gestellt werden kann,

c) Rückschlüsse auf größere Zusammenhänge sprachlicher Art möglich sind.

3. Es darf nichts von vorn herein ausgeschlossen werden, was sprachwissenschaftlich interessant werden könnte.

4. Die Texte sollen aus möglichst in sich geschlossenen einzelnen Texteinheiten bestehen.

5. Es darf bei der Aufnahme möglichst kein Informationsverlust eintreten; dazu ist u.a. eine umfassende, genaue Kennzeichnung der einzelnen Texteinheiten erforderlich.

6. Das Material muß insgesamt und in seinen Teilen möglichst schnell und leicht zugänglich und reproduzierbar sein.

Schon im Hinblick auf Punkt 6 erweist sich das Mittel der maschinellen (elektronischen) Datenverarbeitung als unumgänglich.

4.2. Zeitliche Einheiten und Einteilungen

Zeitungen sind, im Gegensatz zu Büchern, kein in sich geschlossener, abgeschlossener Zusammenhang gestalteter Sprache, sondern ein kontinuierlicher, sich in bestimmten, regelmäßigen Abständen (Tagen, Wochen usw.) erneuernder Strom von Sprache.
Markierungen, die zum Ziel haben, innerhalb dieses Stroms "Anfang" und "Ende" zu setzen, haben - abgesehen von der durch Drucktechnik und Erscheinungsweise gegebenen Unterteilung in einzelne Ausgaben - alle etwas Unverbindliches an sich; die gebräuchlichste Einteilung ist jedoch die nach Jahrgängen [13]. In jedem Fall ist die Einheit des Jahrgangs die größte

Einteilung, die sich aus den Zeitungen selbst gewinnen läßt[14]; es ist kein allgemeines Kriterium erkennbar, nach welchem man mehrere Jahrgänge zu einer Einheit zusammenfassen könnte.

Andere Einteilungen nach kleineren Einheiten, etwa nach Quartalen, Monaten oder Wochen, sind denkbar. Welche Einheit oder Einheiten man zur Grundlage der Dokumentation wählt, hängt ausschließlich von den Fragestellungen ab, die man an das Material zu stellen beabsichtigt, bzw. von dem Zweck, für den es bestimmt ist.

Grundsätzlich muß gelten: Je allgemeiner oder vielschichtiger die schon formulierten oder zu erwartenden Fragestellungen sein können, um so ausgedehnter muß das Material und der Zeitraum sein, aus dem es stammt; je enger und spezieller die zu erwartenden Fragestellungen, um so kleiner darf der Zeitraum gewählt werden, um so stärker werden dann aber auch zeitraumbedingte und damit auch themenbedingte Eigenheiten des Materials, besonders im Wortschatz, hervortreten. Der Bearbeiter muß diese Eigenheiten im voraus berücksichtigen; er hat sich darüber Rechenschaft abzulegen, warum er gerade das Material aus d i e s e m Zeitraum (z.B. Woche) und nicht aus einem der vielen möglichen anderen gewählt hat.

Für allgemeine Erörterungen und die sehr vielschichtigen Fragestellungen, mit denen wir rechnen müssen, ist zweckmäßigerweise von der G r o ß e i n h e i t d e s J a h r g a n g s auszugehen.

4.3. Mengenproblem und Kapazität

Vor Beginn der Textaufnahme muß Klarheit bestehen über die Dimensionen, in denen wir uns bei der Aufnahme von Zeitungstexten bewegen. Die WELT hat im Jahrgang 1964 (ohne Beilagen!) in ca. 310 erschienenen Nummern ca. 5.600 Zeitungsseiten mit einer Textmenge von rund 15 Millionen laufen-

den Wörtern produziert. Das entspricht einem Buchwerk von über 30.000 Seiten. Eine geübte Schreibkraft brauchte etwa 5.500 Arbeitstage entspr. 24 Jahren (bei rund 230 Arbeitstagen p.a.), um diese Menge auf Datenträger zu übertragen; sie brauchte etwa die gleiche Zeit für die Korrektur[15]. Selbst die von einer weniger umfangreichen Zeitung, etwa dem "Neuen Deutschland", in nur einem Monat produzierte Textmenge entspricht noch immer dem Umfang eines Buches von über 1.200 Seiten[16].

Es ist offensichtlich, daß solche Mengen die Schreibkapazität der weitaus meisten Institute überschreiten. Solange die elektronische Industrie nicht ein auch für Zeitungen brauchbares optisches Lesegerät auf den Markt bringt, wird die Aufnahme vollständiger Zeitungsjahrgänge also nicht möglich sein. Auch die Aufnahme ganzer Monats-Mengen ist noch sehr aufwendig. Hier und um so mehr bei noch kleineren Einheiten, die vollständig zu bewältigen wären, stellen sich jedoch die unter Absatz 4.2. erwähnten Bedenken ein.

Theoretisch besteht noch die Möglichkeit, sich die Tatsache zunutze zu machen, daß die meisten Zeitungsdruckereien ihre Setzmaschinen mittels Lochstreifen (TTS-Streifen) steuern. Sofern man sich regelmäßig diese Lochstreifen beschaffen kann, wäre es möglich, auch sehr große Textmengen ohne eigene Schreibarbeit elektronisch zu speichern. Leider war diese Möglichkeit wegen einer Reihe praktischer Schwierigkeiten bisher nicht zu verwirklichen[17], abgesehen davon, daß mittels TTS-Streifen ohnehin kein älteres Material zu gewinnen ist.

Aus diesen Gründen erweist es sich als unvermeidlich, eine Auswahl aus dem jeweils vorgesehenen Gesamtmaterial zu treffen.

4.4. Begründbarkeit einer Textauswahl

Eine vollständige Aufnahme eines umfangreichen Textkorpus wäre aber zudem nur dann erforderlich, wenn es darum ginge, die Eigenheiten einzelner Teile dieses Korpus (etwa eines Romanwerkes, Zeitungsjahrgangs usw.) vollständig zu erforschen, d.h. wenn Fragestellungen zu erwarten sind, die sich auf dieses eine Werk in seiner Einmaligkeit richten. Dies ist aber im Rahmen der Dokumentation, die das Institut einschließlich der Außenstelle Bonn treibt bzw. zu treiben beabsichtigt, durchweg nicht der Fall. Die aufgenommenen Texte sollen nur Beispiele sein für die deutsche Sprache (östliche, westliche Zeitungssprache, Literatursprache, Trivialliteratursprache usw.) der Gegenwart. Ein "repräsentatives" Gesamtkorpus wird sich also erst aus einer gut zusammengestellten Mischung verschiedener solcher Texte zusammensetzen. Dann aber erfüllt eine Auswahl aus vielen solcher Texte ihren Zweck besser als wenige vollständig aufgenommene Texte, auch und gerade dann, wenn neben synchronischen auch diachronische Fragestellungen zu erwarten sind. Dies gilt ganz besonders für Untersuchungen zum Wortschatz (jedoch nicht nur für diese), da erwiesen ist, daß bestimmte Eigentümlichkeiten des Wortschatzes großer Textkorpora schon in Teilmengen gut erkennbar sind (Homogenität vorausgesetzt) [18].

Damit kann das Prinzip der Auswahl nicht nur aus praktischen Gründen notwendig, sondern auch als sinnvoll, im Hinblick auf die höhere Arbeitseffektivität im Vergleich zum Prinzip der vollständigen Textaufnahme sogar als besser bezeichnet werden.

Es ist somit zu unterscheiden zwischen der Gesamtmenge, d.h. derjenigen Menge, über die letzlich Aussagen gemacht werden sollen, und der tatsächlich aufgenommenen Auswahlmenge (Modellmenge), d.h. derjenigen Menge, mittels derer Aussagen über die Gesamtmenge gemacht werden sollen.

Ein wesentlicher Nachteil liegt zweifellos darin, daß es nicht möglich ist, experimentell zu beweisen, daß die Auswahl tatsächlich Abbild der Gesamtmenge ist, da diese ja für genaue Textanalysen gerade nicht zur Verfügung steht. Über den Grad der Entsprechung sind nur Wahrscheinlichkeitsaussagen möglich. Um so sorgfältiger muß die Aufnahme der Auswahl vorbereitet werden.

4.5. Bedingungen einer zureichenden Auswahl [19)]

4.5.1. Allgemeine Feststellungen

Um eine Auswahl treffen zu können, sind folgende Schritte erforderlich:

1. Es ist zunächst die Gesamtmenge zu bestimmen und zu sichten, d.i. die Großeinheit von Texten, auf die sich die Dokumentation beziehen soll.
2. Innerhalb dieser (Gesamt-) Bezugsmenge ist die Zielmenge zu bestimmen, d.h. die Menge, aus der die aufzunehmende Auswahl gewonnen wird.
3. Die Zielmenge muß in allen ihren Eigentümlichkeiten so genau wie möglich bestimmt werden.
4. Aus dieser Zielmenge muß eine Auswahl gewonnen werden, die
   a) alle Eigenheiten der Zielmenge möglichst maßstabgetreu widerspiegelt,
   b) ihre Menge nach ausreicht, um statistische Untersuchungen sowie Rückschlüsse auf die Zielmenge zu erlauben;
   c) alle die auf Seite 47/48 aufgeführten sechs Anforderungen, die für jede Dokumentation gelten, erfüllt,

d.h. es muß ein qualitativ und quantitativ zureichendes Modell erstellt werden[20)].

Erläuterungen

Zu 1:

Es ist für die zu schaffende Auswahl von großer Bedeutung, ob sie sich etwa auf ein einzelnes Werk, das Gesamtwerk eines Dichters, eine literarische Gattung oder Epoche usw. beziehen soll. Je komplexer oder ausgedehnter die bezogene Gesamtmenge ist, um so komplexer oder ausgedehnter muß auch die Auswahl insgesamt und in ihren Teilen sein.

Zu 2:

Die Unterscheidung von Gesamtmenge und Zielmenge ist erforderlich, weil nicht immer alle Teile eines großen Textkorpus, etwa des Gesamtwerkes eines Schriftstellers, eines gesamten Zeitungsjahrgangs samt allen Beilagen usw., in die Dokumentation einbezogen werden sollen. Es kann sinnvoll sein, bestimmte Teile, wie die Briefe, Tagebücher und Entwürfe aus dem Gesamtwerk eines Schriftstellers bzw. die literarische Beilage aus einem Zeitungsjahrgang, zu übergehen und eine entsprechend eingeschränkte Zielmenge zu definieren[21].

Zu 3:

Zu den Eigentümlichkeiten der Zeitung gehört zum Beispiel, daß ihr Textmaterial nicht nur eine "vertikale" Ausdehnung hat - meßbar an der jeweils zu einem bestimmten Zeitpunkt erschienenen Zahl von Druckseiten -, sondern auch eine "horizontale" - meßbar an der Zahl der in einem bestimmten Zeitraum in einem bestimmten Rhythmus erschienenen Ausgaben[22]. Verbunden mit den auf S.42/43 erwähnten grundsätzlichen Eigentümlichkeiten, besonders der Verfasser-, Themen- und Formenvielfalt, erweist sich die Zeitung also als ein äußerst komplexes Gebilde. Verfahren zur Dokumentation etwa bei Büchern oder auch weniger kompliziert aufgebauten Periodika können wahrscheinlich im wesentlichen als Vereinfachungen des bei Zeitungstexten anzuwendenden Verfahrens betrachtet werden.

4.5.2. Qualitative Analyse

4.5.2.1. Feststellung spezifischer Dominanzen

Wenn angenommen werden könnte, daß die verschiedenen Eigenheiten (wie Einfluß bestimmter Verfassergruppen, Themenwahl, Sachgebiete, Berichtsformen usw.) gleichmäßig über die ganze Jahrgangsmenge in vertikaler und horizontaler Richtung verteilt wären, wäre die Gewinnung einer Modellmenge kein Problem; man könnte sich darauf beschränken, in regelmäßigen Abständen etwa gleich große Textmengen an beliebiger Stelle zu entnehmen und sie zu einer Auswahlmenge zu vereinigen. Leider verhält es sich nicht so; vielmehr zeigen alle Zeitungen in jeder Ausgabe gewisse regelmäßige Dominanzen, die es unmöglich machen, die Auswahl dem Zufall zu überlassen, sofern Verzerrungen und Einseitigkeiten vermieden werden sollen.

a) Dominanzen in vertikaler Richtung:

Die meisten Zeitungen gliedern ihre Ausgaben in S p a r t e n . Im Regelfalle sind die ersten 2 - 3 Seiten der Politik und aktuellen Nachrichten vorbehalten, darauf folgt "Wirtschaft" und "Kulturelles" (Feuilleton). Die letzte Seite enthält zumeist Vermischtes, dazwischen können Lokalseiten oder Seiten für Sport, Reise usw. eingeschaltet sein. Gelegentlich können bestimmte Sparten zu Lasten anderer stark erweitert werden. - Sofern eine Auswahl nicht absichtlich bestimmte Sparten und damit Sachgebiete und Themen bevorzugen und andere vernachlässigen will, muß auf eine gleichmäßige Berücksichtigung aller Seiten geachtet werden, d.h. alle Seiten von der ersten bis zur letzten müssen entsprechend ihrer Verteilung in der Zielmenge auch in der Auswahlmenge enthalten sein.

b) Dominanzen in horizontaler Richtung:

(Wochenturnus)

Es läßt sich feststellen, daß bestimmte Themen und Sparten sich an bestimmten Tagen häufen. Montags nimmt durchweg der Sport einen breiten Raum ein, wogegen Wirtschaft, vor allem Börsenberichte (in westlichen Zeitungen), weitgehend fehlen. Zum Wochende hin verstärkt sich der Anteil von Artikeln zum Thema "Reise, Erholung, Freizeit" usw., samstags, oft auch mittwochs, häufen sich Kleinanzeigen, auch Feuilleton und Unterhaltung. Eine bestimmte Wochentagsausgabe enthält meist Berichte über Wissenschaft, Literatur usw. Auch Leserbriefe werden oft nur in bestimmten Wochentagsausgaben gebracht.
Es besteht ferner Grund zu der Annahme, daß bestimmte Verfasser bevorzugt an bestimmten Wochentagen Artikel veröffentlichen.

Auch hier kann es also nicht gleichgültig sein, ob bestimmte Wochentagsausgaben häufiger als andere in der Modellmenge enthalten sind. Es ist wiederum auf gleichmäßige Berücksichtigung a l l e r Wochentagsausgaben zu achten.

(Monatsturnus)

Regelmäßigkeiten im Ablauf der einzelnen Monate ließen sich bisher nicht erkennen.

(Quartalsturnus)

Regelmäßigkeiten innerhalb der Quartale sind nur sehr schwach und nur im Wirtschaftsteil zu erkennen: Jeweils beim Übergang zu einem neuen Quartal finden sich statistische Übersichten über Wirtschaftsentwicklungen häufiger als sonst.

(Jahresturnus)

Innerhalb eines Jahrgangs sind deutlich ausgeprägte Dominanzen zu erkennen. Der Wechsel von Saat und Ernte prägt sich besonders in östlichen Zeitungen im Bereich der Wirtschaft außerordentlich stark aus; Jahresende und -beginn bringen eine ungewöhnliche Häufung von Übersichten, Entwicklungsberichten, Prognosen, Planungen usw., desgleichen die Zeit der Beratung und Verabschiedung des Haushalts (Frühjahr bzw. Herbst). Die Sommerpause wirkt sich auf den Sportteil (bes. Fußball) ebenso negativ wie auf die Sparte "Reise, Urlaub" usw. positiv aus. Die saisonbedingte Dominanz bestimmter Sportarten (Wintersport, Wassersport) ist ebenso spürbar wie der Einfluß der großen Feste (Weihnachten, Karneval, Ostern), und zwar nicht nur im Werbe-Teil. Unabhängig von diesen jährlich wiederkehrenden Dominanzen können selbstverständlich bestimmte Ereignisse (Olympiade, Nahostkrieg u.ä.) die Berichte der entsprechenden Sparte über längere Zeit hin nachhaltig beeinflussen.

Jede Abweichung von einer gleichmäßig über den ganzen Jahrgang hin verteilten Auswahl kann daher zu erheblichen Verschiebungen vor allem im Wortschatz führen. Die Bevorzugung etwa eines bestimmten Monats in einer Auswahl macht diese jedenfalls als Modellmenge für den Jahrgang unbrauchbar. Gleiches gilt für die Bevorzugung bestimmter Seiten, Sparten oder Artikelformen.

Die hier aufgezeigten Regelmäßigkeiten gelten nur für eine bestimmte Art von Tageszeitungen und auch für diese nicht ohne Einschränkung. Abweichungen sind jederzeit möglich. Für andere Arten von Tageszeitungen, etwa für Boulevardblätter[23)], müssen neue Untersuchungen angestellt werden. Es ist außerdem möglich, daß wesentliche Eigenheiten übersehen werden, die sich dann unkontrolliert verzerrend auf die in der Modellmenge herr-

schenden Verhältnisse auswirken[24]. Andererseits ist es möglich, Kriterien aufzugreifen und bei der Auswahl zu berücksichtigen, die sich als irrelevant erweisen können[25].

4.5.3. Quantitative Analyse

4.5.3.1. Definition der Mengeneinheiten

Bevor man nach der Feststellung der qualitativen Eigenheiten der Zielmenge zur Ermittlung der Modellmenge übergehen kann, ist es notwendig, die Mengeneinheiten in der Zielmenge festzustellen bzw. zu definieren. Als Meßeinheit[26] bietet sich die S e i t e als einwandfrei formal definierbare Einheit an, die im übrigen den Vorteil hat, daß sie sowohl für Texte rein vertikaler Ausdehnung (Bücher) als auch für Texte vertikaler und horizontaler Ausdehnung (Periodika) gilt[27].

a) Vertikale Mengeneinheiten

Zeitungen erscheinen jedoch nicht in "Seiten", sondern in "Ausgaben". Der Begriff "Ausgabe" ist jedoch unklar. Es fragt sich, ob etwa Beilagen dazu gehören oder nur bestimmte Beilagen[28]. Tatsächlich kann sich die Menge, mit der eine Zeitung erscheint, zusammensetzen aus

1. der Grundmenge (gewöhnlich der einheitlich durchnumerierte Teil der Zeitung)
2. den regelmäßigen Beilagen (meist gesondert numeriert)
3. den unregelmäßigen Beilagen, den Sonderbeilagen zu einmaligen Ereignissen, Extrablättern usw. (gesondert oder nicht numeriert).

Die unter Einschluß aller Erweiterungen und Beilagen an einem Tag erschienene Menge nenne ich E r s c h e i n u n g s m e n g e . Der Bearbeiter hat sich also zunächst zu entscheiden, ob er die gesamte Erscheinungsmenge oder nur einen Teil von ihr - und welchen - zur Grundlage

seiner weiteren Berechnungen machen will. Die Mengeneinheit, für die er sich entscheidet, nenne ich Ausgabe [29]. Erscheinungsmenge $S^{(e)}$ und Ausgabe $S^{(a)}$ sind in Zahl der Seiten (S) meßbar, wobei S_e eine aus mehreren Teilen (G=Grundmenge, E1 und E2 = Erweiterungen 1 und 2) zusammengesetzte Größe ist:

$$S_G + S_{E1} + S_{E2} = S_e \; ; \text{ dabei ist } S_e \gtreqless S_a$$

Umfang der Ausgabe wie der Erscheinungsmenge können täglich wechseln. Die Erscheinungsmenge unterscheidet sich von der Ausgabe also in gleicher Weise wie die Gesamtmenge von der Zielmenge.

b) Horizontale Mengeneinheiten

Im folgenden wird vorausgesetzt

1. daß die Zielmenge 1 Jahrgang ist,
2. daß die Ausgabe die gesamte Erscheinungsmenge umfaßt (also $S_a = S_e$).
3. daß sich innerhalb des Jahrsgangs die Erscheinungsweise der Zeitung nicht grundlegend ändert [30].

Zur Ermittlung der horizontalen Mengenverteilung hat der Bearbeiter eine Reihe von Feststellungen zu treffen:

1. Feststellung der Zahl der Ausgaben (A_J) pro Jahrgang (J), wobei die Zahl der Ausgaben gleich der Nummer (n) der letzten Ausgabe des Jahres ist: $A_J = n$
2. Feststellung der Gesamtzahl der im Jahrgang erschienenen Seiten:

   $S_{a1} + S_{a2} + S_{a3} + \ldots + S_{an} = S_{aJ}$

   (Daraus ergibt sich die durchschnittliche Seitenzahl der Ausgaben $s_a = \frac{S_{aJ}}{n}$)

3. Feststellung des Erscheinungsintervalls (des in Tagen zu messenden Abstands zwischen den Ausgaben)

   a) Das durchschnittliche Erscheinungsintervall läßt sich errechnen als $\frac{365}{n}$, wobei sich bei täglich erscheinenden Ausgaben ein Wert nahe 1, bei Tageszeitungen ohne Sonntagsausgabe ein Wert nahe 1,2 und bei Wochenzeitungen nahe 7 ergeben muß [31].

   Geringe Abweichungen nach oben sind normal, da immer einige Zeitungsausgaben, abweichend vom regelmäßigen Erscheinungsintervall, ausfallen (wegen Betriebsstörungen, an Festtagen usw.)[32].

4.5.4. Feststellen der Seitenhäufigkeiten

Da die Ausgaben einer Zeitung verschieden lang sind, erscheinen nicht alle Seiten gleich häufig in der Zielmenge. Da aber bestimmte Sparten und damit Themen an bestimmte Seiten oder Seitengruppen mehr oder weniger eng gebunden sind, ist es notwendig, die Verteilung der einzelnen Seiten in der Zielmenge zu ermitteln, um sie in der Modellmenge berücksichtigen zu können. Bei der Aufnahme einer Zeitung, deren Umfang zwischen 8 und 20 Seiten schwankt, muß die Auswahl mehr Seiten der Seitennummern 1 - 8 als der Seitennummern 12 - 20 enthalten.

4.5.4.1. Seitenhäufigkeiten in der Zielmenge

a) Durch eine teilweise [33] Auszählung des Jahrgangs ist der Umfang der einzelnen Ausgaben festzustellen.
   Mindestumfang der Ausgaben: y Seiten
   Maximalumfang der Ausgaben: y + m Seiten
   (m = Differenz zwischen der kleinsten und der größten Ausgabe in Seiten)

b) Verteilung der Seiten nach Ausgaben verschiedener Länge (Klassenbildung):

p	Ausgaben	enden	mit	Seite	y	,	enthalten	also	p mal y Seiten
p_1	"	"	"	"	y + 2,	"	"	p_1	" (y + 2) "
p_2	"	"	"	"	y + 4,	"	"	p_2	" (y + 4) "
p_3	"	"	"	"	y + 6,	"	"	p_3	" (y + 6) "
.					.				
.					.				
.					.				
$\frac{p_i}{n}$	"	"	"	"	y + m,	"	"	p_i	" (y + m) "
								S_{aJ}	

c) Ermittlung der Häufigkeit jeder Seite pro Klasse in der Zielmenge:

Es kommen im Material vor:

die Seiten 1 bis y in jeder Ausgabe, also je n-mal; die Seiten y + 1 und y + 2 je n - p mal usw.:

S. 1	n	⎫
S. 2	n	⎪
S. 3	n	⎪
. .	.	⎬ zus. y mal n
. .	.	⎪
. .	.	⎪
S. y	n	⎭
S. y + 1	n - p	
S. y + 2	n - p	
S. y + 3	$n - (p + p_1)$	
S. y + 4	$n - (p + p_1)$	
S. y + 5	$n - (p + p_1 + p_2)$	
S. y + 6	$n - (p + p_1 + p_2)$	
. .	.	.
. .	.	.
. .	.	.
. .	.	.

$S.\ y + m - 1 \qquad n - (p + p_1 + p_2 + \ldots + p_i)$

$S.\ y + m \qquad n - (p + p_1 + p_2 + \ldots + p_i)$

Den jeweils gefundenen Wert nenne ich F.

4.5.4.2. Seitenhäufigkeiten in der Auswahlmenge

(Umrechnung auf die Anteile der einzelnen Seiten in der Auswahlmenge)

Die gesuchte Häufigkeit f jeder Seite pro Klasse in der Auswahlmenge (Modellmenge) M verhält sich zur Gesamtzahl der Seiten (S_M) in der Auswahlmenge wie die Häufigkeit F jeder Seite pro Klasse in der Zielmenge zur Gesamtzahl der Seiten (S_{aJ}) in der Zielmenge:

$$\frac{f}{S_{M.}} = \frac{F}{S_{aJ}} \; ; \qquad f = F. \frac{S_M}{S_{aJ}} \quad , \text{ wobei F jeweils einsetzbar,}$$

$\frac{S_M}{S_{aJ}}$ ein konstanter Faktor ist.

Sollten nicht für die gesamte Zielmenge, sondern nur für einige Teile von ihr exakt ausgezählte Werte vorliegen, so sind zweckmäßigerweise sowohl bei der Ermittlung von F als auch von f die tatsächlich ausgezählten Werte zugrundezulegen[34].

Sofern die Größe der Auswahlmenge bekannt ist, kann damit der Anteil aller Seiten in der Auswahl bestimmt werden.

Die Größe der Auswahlmenge kann vorgegeben sein (etwa durch Kapazitätsgrenzen), sie kann auch abhängen von der Auswahlquote (in Prozent der Zielmenge), vom Wahrscheinlichkeitsquotienten oder von Zahl und Umfang vorgegebener Stichproben.

4.6. Ermittlung der Modellmenge

Es gibt eine ganze Reihe von Möglichkeiten, eine Modellmenge sinnvoll zusammenzustellen. Welche dieser Möglichkeiten gewählt wird, hängt in erster Linie von dem Zweck ab, welcher der Materialaufnahme zugrundeliegt, in zweiter Linie von den Ansprüchen, die an das Material gestellt werden.
Nach diesen Gesichtspunkten sind dann einige Vorentscheidungen zu treffen:

4.6.1. Aufnahmeeinheit

Als erstes ist die Aufnahmeeinheit, die der Modellmenge zugrundeliegt, zu bestimmen. Wenn bisher vorausgesetzt wurde, daß die Seite als formal eindeutige Grund- und Zähleinheit der Zielmenge auch als Aufnahmeeinheit dient, so geschah das aus praktischen Gründen und weil ihr auch eine thematisch-inhaltlich gliedernde Funktion im Aufbau einer Ausgabe zukommt. Außer der Seite ist auch etwa der "Artikel" oder die "Ausgabe" als Aufnahmeeinheit denkbar. Der Artikel bietet den Vorteil sehr fein differenzierbarer Streuung und thematischer Einheitlichkeit, den Nachteil schwieriger Definition und sehr unterschiedlichen Umfangs; die Seite bietet den Vorteil mittlerer Streuungsmöglichkeit, formaler Eindeutigkeit und begrenzter thematischer Gebundenheit, den Nachteil unterschiedlichen Umfangs und den Nachteil, daß Textzusammenhänge zerrissen werden können, da gelegentlich Artikel auf einer Seite begonnen und auf einer anderen fortgesetzt werden. Die Ausgabe hat den Vorteil relativer Geschlossenheit, den Nachteil zu großen Umfangs, der dazu führen kann, daß nur sehr wenige Ausgaben aufgenommen werden können.

4.6.2. Stichprobeneinheit

Wird aus diesem Grund die Ausgabe als Aufnahmeeinheit abgelehnt und eine der beiden kleineren Einheiten gewählt, ist noch zu entscheiden, wie viele von diesen Aufnahmeeinheiten bei jeder Stichprobe jeweils aufgenom-

men werden sollen. Neben der Aufnahmeeinheit ist damit der weitere Begriff der Stichprobeneinheit einzuführen. Sie hat mindestens die Größe der Aufnahmeeinheit oder eines Mehrfachen von dieser [35].

4.6.3. Aufnahmeintervall, Aufnahmedichte

Diese Überlegung führt zur Frage des Aufnahmeintervalls. Das Aufnahmeintervall ist der zeitliche Abstand zwischen zwei Stichproben, und zwar, dem Grundsatz der gleichmäßigen Auswahl entsprechend, ein möglichst gleich großer Abstand zwischen allen Stichproben. Da immer davon ausgegangen werden muß, daß für die Herstellung der Modellmengen nur eine bestimmte, begrenzte Arbeitskapazität zur Verfügung steht, muß das Aufnahmeintervall um so größer, d.h. die Aufnahmedichte um so geringer werden, je größer die einzelnen Stichprobeneinheiten sind. Es gibt jedoch eine obere Grenze[36] für die Größe des Aufnahmeintervalls (= Untergrenze der Aufnahmedichte): sie wird dann überschritten, wenn mit hoher Wahrscheinlichkeit zu befürchten ist, daß sogar wesentliche sprachliche Erscheinungen, soweit sie von horizontalen Dominanzen her beeinflußt sind, von den Stichproben nicht mehr berührt werden[37].

4.6.4. Zahl der Stichproben

Die Zahl der Stichproben in der gesamten Modellmenge ist abhängig vom Aufnahmeintervall oder umgekehrt. Allerdings läßt sich, ebenso wie für das Aufnahmeintervall, eine Grenze angeben:
Die Zahl der Stichproben (meist identisch mit der Zahl der Aufnahmeeinheiten) darf nicht kleiner sein als die Zahl der für den Beobachter wichtigen Kriterien, die in der Auswahl zu berücksichtigen sind, bzw. der Rhythmus des Auftretens dieses wichtigen Kriteriums. Hat etwa eine Zeitung einen durchschnittlichen Umfang von 20 Seiten und soll jede Seite mindestens 1 mal in der Modellmenge enthalten sein, sind mindestens 20 Stichproben (=Seiten) erforderlich. Ist der Wochenrhythmus ein wich-

tiges Kriterium, sind 52 Stichproben erforderlich; sind 100 verschiedene Sachgebiete in einer Zeitung festgestellt und soll jedes fünfmal vertreten sein, benötige ich mindestens 500 Artikel usw. Eine Mindestmenge läßt sich also meist schon von der Fragestellung her bestimmen.

4.6.5. Wahrscheinlichkeitsquotient

Ob diese Menge tatsächlich ausreicht, d.h. ob eine qualitativ zureichend ermittelte Auswahl auch rein quantitativ die Bedingungen einer aussagekräftigen Modellmenge erfüllt, ist eine mathematisch-statistische Frage, die im Teil III des Forschungsberichts mit entsprechenden Methoden untersucht wird. Das Ergebnis dieser Berechnungen ist der Wahrscheinlichkeitsquotient, der ausdrückt, mit welcher Wahrscheinlichkeit eine in der Zielmenge vorhandene seltene Erscheinung (Wortform) auch in der Modellmenge erwartet werden kann und umgekehrt[38].

4.6.6. Aufnahmezeitraum

Der Zeitraum, aus dem die Stichproben genommen werden, muß genau dem Zeitraum entsprechen, für den die Modellmenge aussagekräftig sein soll. Jede Auswahl kann Modell sein nur für den Zeitraum, der durch die erste und die letzte Aufnahmeeinheit - plus je einem halben Aufnahmeintervall vor der ersten und nach der letzten Aufnahmeeinheit - begrenzt wird[39].

4.7. Zusammenfassung

der erforderlichen Schritte zur Ermittlung einer Modellmenge

Unter der Voraussetzung, daß

a) die Rahmen-Zielsetzung der Dokumentation,

b) der Rahmen der Kapazität (Arbeitskräfte, Hilfsmittel, Zeit)

festliegen, sind folgende Schritte notwendig:

1. Festlegung der Gesamtmenge und Erscheinungsmenge
   daraus: Festlegung der Zielmenge und der Ausgabe
2. Bestimmung der Eigentümlichkeiten der Zielmenge
   a) Dominanzen in vertikaler Richtung
   b) Dominanzen in horizontaler Richtung
3. Messung (Auszählung) der Zielmenge,
   der Ausgaben und
   des Erscheinungsintervalls
   in Seiten bzw. Tagen
4. Feststellung der Seitenverteilung in der Zielmenge
   daraus: Ermittlung der Seitenanteile (F)
5. Bestimmung der Aufnahmeeinheit für die Auswahlmenge
6. Bestimmung des Aufnahmeintervalls (→ Aufnahmedichte)
   daraus: Ermittlung der Zahl der Stichproben
7. Bestimmung der Aufnahmequote nach dem gewünschten Wahrscheinlichkeitsquotienten (bzw. umgekehrt)
8. Errechnung der insgesamt aufzunehmenden Auswahlmenge in Aufnahmeeinheiten (sofern nicht vorgegeben), damit:
   Bestimmung des Umfangs der Stichproben
9. Errechnung der Seitenverteilung f in der Auswahlmenge (unter Berücksichtigung von Punkt 4)
10. Verteilung der Aufnahmeeinheiten (Seiten) auf die Stichproben und der Stichproben über den ganzen Aufnahmezeitraum anhand des unter 9 ermittelten Seitenschlüssels.
    (Bei diesem Arbeitsvorgang ist, zur Vermeidung des Vorwurfs willkürlicher Verteilung, ein möglichst fester Verteilungsturnus anzustreben.)

Je nachdem, ob die Aufnahmequote, Umfang und Verteilung der Auswahlmenge oder der Wahrscheinlichkeitsquotient vorgegeben oder abgeleitet sind, muß die Reihenfolge der Schritte 7 - 10 entsprechend vertauscht werden.

Spezieller Teil : Dokumentation und maschinelle Verarbeitung der Bonner Zeitungsjahrgänge

5. Allgemeines

Zu den Zielen und Methoden der in der Außenstelle betriebenen Dokumentation bei Zeitungstexten siehe die allgemeinen Bemerkungen unter Punkt 1 bis 3 und 4.1. bis 4.2. Entsprechend den dort zusammengestellten grundsätzlichen Bemerkungen sollte sich die Dokumentation auf mehrere ganze Jahrgänge überregionaler Tageszeitungen politischen Charakters und "repräsentativer" Art aus West und Ost beziehen. Da östliches Zeitungsmaterial schwerer zugänglich ist als westliches und da sich zudem das wissenschaftliche Interesse zunächst mehr auf das östliche Material richtete, sollte mit einem Zeitungsjahrgang aus der DDR begonnen werden.

6. Zur Kapazität

Für die Übertragung der Texte auf Datenträger standen der Außenstelle zwei bis vier, durchschnittlich 3,5 Halbtagskräfte zur Verfügung, von denen jeweils zwei mit dem Abschreiben, die übrigen mit der Korrektur und dem Lesen der abgeschriebenen Texte beschäftigt waren [40]. Es mußte also von einer Schreibkapazität von 2 mal 20 Wochenstunden, also (bei durchschnittlich 46 Arbeitswochen pro Jahr) von 1840 Arbeitsstunden pro Jahr ausgegangen werden.

Da zum Abschreiben einer Zeitungsseite, sofern alle Schreibkonventionen [41] beachtet werden, ungefähr 9 Arbeitsstunden benötigt werden, ergibt sich eine theoretische Jahreskapazität von etwas über 200 Seiten. Durch den gerade bei Zeitungstexten sehr hohen Aufwand an Korrekturarbeiten sinkt die Schreibkapazität um rund ein Viertel, da etwa dieser Anteil der Seiten nach beendetem Korrekturen-Lesen neu geschrieben werden muß. Die obere Grenze der Kapazität lag also bei 150 Seiten pro Jahr [42].

Wir gingen bei der Arbeitsplanung davon aus, daß grundsätzlich mehr als ein Jahrgang, also mindestens zwei, pro Jahr abgeschrieben und zur Verfügung gestellt werden sollten. Daraus ergab sich, daß die aus jedem Jahrgang aufgenommene Menge im Durchschnitt nicht mehr als 70 bis 80 Seiten umfassen durfte.

7. Bestimmung der Gesamtmengen

Als Grundlage für die Dokumentation von Zeitungstexten aus der DDR wurde

1. der Jahrgang 1964 des "Neuen Deutschland" (Organ des Zentralkomitees der SED), Berliner Ausgabe[43],

für Zeitungstexte aus der Bundesrepublik

2. der Jahrgang 1964 der WELT, Hamburg, Berliner Ausgabe,

gewählt.

Für vergleichende Untersuchungen diachronischer Art wurde als zweite Stufe der Textaufnahme

3. der Jahrgang 1954 des "Neuen Deutschland"[44],

und

4. der Jahrgang 1954 der WELT, Berliner Ausgabe,

ausgewählt.

7.1. Begründung zur Wahl der Ost-Jahrgänge

Das "Neue Deutschland" ist im Rahmen der von Staat und Partei kontrollierten Presse der DDR die einzige überregionale, in der ganzen DDR verbreitete Tageszeitung (im Jahre 1964 täglich, auch sonntags, erscheinend). Sie ist als Zentralorgan der SED in besonderer Weise repräsentativ für die

Meinung (und Sprache) der in der DDR herrschenden Gruppe und maßgebend für den größten Teil der DDR-Presse. Das "Neue Deutschland" (ND) ist nicht in erster Linie Verkaufsobjekt, sondern bewußt Erziehungs-, Propaganda- und Führungsmittel der Partei.

In der Erfüllung dieser Funktion ist es extrem "politisch" orientiert, d.h. es schränkt die Vielfalt und den Umfang der aus westlichen Zeitungen gewohnten Themen und Sparten erheblich ein bzw. bezieht diese mit in den Bereich des Politischen ein[45]. Mit dem Jahrgang 1964 wurde deshalb begonnen, weil dieser der letzte war, der bei Beginn der Arbeiten (Anfang 1965) abgeschlossen vorlag. In der zweiten Stufe (Jahrgang 1954) wurde ein Zehnjahresabstand gewählt, da angenommen wird, daß ein relativ großer Abstand nötig ist, um wesentliche Veränderungen im Wortschatz in West und Ost sichtbar zu machen. Es ist jedoch geplant, den Abstand später zu halbieren und noch je einen Jahrgang 1959 aufzunehmen.

7.2. Begründung zur Wahl der West-Jahrgänge

Für den Bereich der Bundesrepublik läßt sich keine Tageszeitung benennen, die in entsprechender Weise repräsentativ genannt werden kann. Die auflagenstärkste Zeitung, die Bild-Zeitung, ist ein überwiegend unpolitisches Boulevard-Blatt; unter den überregionalen "seriösen" Zeitungen, der "Frankfurter Allgemeinen Zeitung", der "WELT", vielleicht noch der "Süddeutschen Zeitung", "Stuttgarter Zeitung" und der "Frankfurter Rundschau", eine im gleichen Sinne "repräsentative"zu benennen, ist kaum möglich. Keine von ihnen gibt in dem Maße die Meinung e i n e r Gruppe wieder, wie dies bei dem ND der Fall ist, keine ist auch in dem Maße "regierungsoffiziös". Am ehesten hat die Wahl zwischen FAZ und WELT zu fallen.

Die Wahl fiel schließlich, trotz etwas geringerer Auflage[46], auf die WELT, da sie die wichtigste politische Zeitung des größten deutschen Zeitungskonzerns ist. Auch hier wurde die Berliner Ausgabe herangezogen. Es ist möglich

und geplant, auf die in Mannheim aufgenommenen [47] Texte aus der FAZ zurückzugreifen und sie mit den WELT-Texten zu vergleichen. Sollten sich dabei erhebliche Abweichungen im Wortschatz ergeben, könnten auch FAZ-Texte laufend aufgenommen werden.

7.3. Vergleichbarkeit der ausgewählten Zeitungen

Es leuchtet ein, daß die beiden ausgewählten Zeitungen nach Zielsetzung, Herkunft, Themenwahl, Darstellungsweise usw. nur bedingt miteinander vergleichbar sind. Im strengen Sinne ist aber keine einzelne westdeutsche Zeitung mit dem ND voll vergleichbar. Als Organ der SED entsprächen dem ND am ehesten die Zeitungen der großen Parteien in der Bundesrepublik, in anderer Hinsicht wäre es vielleicht mit einem Gewerkschaftsorgan vergleichbar, in wieder anderer, z.B. in stilistischer Hinsicht eher mit der Bild-Zeitung oder der Deutschen National-Zeitung oder auch, was die Themenwahl betrifft, mit manchen kleineren Regionalblättern.

Es kann aber nicht Aufgabe unserer Dokumentation sein, die erheblichen Unterschiede zwischen einer repräsentativen Zeitung der DDR und der Bundesrepublik künstlich zu überdecken, etwa indem man aus verschiedenen westdeutschen Zeitungen einen Mischtext herstellt, der der Eigenart des ND möglichst nahe käme [48]. Die Eigenart der Textgrundlage ist zu wahren. Zwar dient die Dokumentation nicht letztlich dem Ziel, die Eigenart zweier Zeitungen miteinander vergleichbar zu machen (das wäre eine zeitungswissenschaftliche Fragestellung), aber sie bedient sich der Texte zunächst zweier, später mehrerer Zeitungen mit jeweils besonderen Eigenheiten, um daran die Eigenheiten der publizierten Sprache in beiden Teilen Deutschlands und ihre Entwicklung vergleichend erkennbar zu machen.

8. Bestimmung der Zielmengen

8.1. Abgrenzung gegenüber Gesamtmengen

a) Neues Deutschland 1964:

Als Zielmenge für den gesamten Jahrgang wurde festgelegt:

die zwischen dem 1. Januar und 31. Dezember 1964 produzierte Menge,
mit Ausnahme der Sonntagsausgabe
und der wöchentlich erscheinenden Beilage
"Die gebildete Nation".

b) DIE WELT 1964:

Als Zielmenge wurde festgelegt:

die zwischen dem 1. Januar und 31. Dezember 1964 produzierte Menge,
mit Ausnahme aller gesondert oder römisch numerierten Beilagen,
nämlich:
"Die Welt der Literatur" (14-tägig)
"Die Reise-Welt" (14-tägig)
"Die geistige Welt" (wöchentlich)
sowie der Stellenanzeigen-Beilage in der Samstagsausgabe (wöchentlich).

c) Neues Deutschland 1954:

Als Zielmenge wurde festgelegt:

die zwischen dem 1. Januar und 31. Dezember 1954 produzierte Menge,
und zwar einschließlich der Sonntagsausgabe (anstelle der fehlenden Montagsausgabe),
mit Ausnahme der Unterhaltungsbeilage (14-tägig).

d) DIE WELT 1954:

Als Zielmenge wurde die zwischen dem 1. Januar und 31. Dezember 1954 produzierte Menge festgelegt,

mit Ausnahme der gesondert oder römisch numerierten Beilage (näheres noch nicht ermittelt).

8.2. Quantitative Beschreibung der Jahrgänge

a) für ND 1964:

Zahl der Erscheinungsmengen (e): 360

Gesamt-Seitenzahl (S_{eJ}): 2820

Zahl der Ausgaben (a): 308

Gesamt-Seitenzahl (S_{aJ}): 2125

Durchschnittliche Seitenzahl (s_a): 6.9

Durchschnittliches Erscheinungsintervall:

(bezogen auf Ausgaben):	1,1883
(bezogen auf Erscheinungsmengen):	1,0167

b) für WELT 1964:

Zahl der Erscheinungsmengen (e): 304

Gesamt-Seitenzahl (S_{eJ}): 8550

Zahl der Ausgaben (a): 304

Gesamt-Seitenzahl (S_{aJ}): 5635

Durchschnittliche Seitenzahl (s_a): 18.54

Durchschnittliches Erscheinungsintervall:

(bezogen auf Ausgaben):	1,2038
(bezogen auf Erscheinungsmengen):	1,2038

c) für ND 1954 :

Zahl der Erscheinungsmengen (e) : 304

Gesamt-Seitenzahl (S_{eJ}) : 2163

Zahl der Ausgaben (a) : 304

Gesamt-Seitenzahl (S_{aJ}) : 1976

Durchschnittliche Seitenzahl (s_a) : 6.5

Durchschnittliches Erscheinungsintervall :

(bezogen auf Ausgaben) :	1,2006
(bezogen auf Erscheinungsmengen) :	1,2006

d) für WELT 1954 :

noch nicht ermittelt.

9. Dominanzen

Zu den Dominanzen vgl. das unter Punkt 4.5.2.1. Gesagte. Für das ND wäre noch hervorzuheben, daß die jahreszeitlich bedingten Dominanzen in der Themenwahl dort besonders ausgeprägt sind. Themen zur Saatzeit, zur Getreide-, Kartoffel- und Rübenernte, zur Planerfüllung (um die Zeit des Jahreswechsels) beherrschen oft wochenlang sogar die erste Seite; sie nehmen den Charakter von Propagandakampagnen an. Westliche Zeitungen wie die WELT vermeiden derartige Häufungen selbstverständlich mit Rücksicht auf die Leser.

10. Auswahlmodus

10.1. Auswahlmodus für ND 64 und 54 :

a) Aufnahmeeinheit :

Entsprechend den unter Punkt 4.6.1. dargelegten Überlegungen wurde die S e i t e als Aufnahmeeinheit gewählt. Der Nachteil, daß Textzusammenhänge (Artikel) gelegentlich auseinandergerissen werden, wurde dabei in Kauf ge-

nommen. Sie lassen sich außerdem ggf. ergänzen[49].

b) Aufnahmeintervall:

Mit Rücksicht auf die begrenzte Kapazität konnte das Aufnahmeintervall nur in der Größenordnung zwischen drei Tagen (= 120 Stichproben) und zwei Wochen (= 26 Stichproben) gewählt werden. Wir entschieden uns für ein durchschnittliches Aufnahmeintervall von einer Woche (= 52 Stichproben).

c) Umfang der Stichproben:

Bei dem gewählten Aufnahmeintervall bestand nur die Wahl zwischen Stichprobeneinheiten von einer oder zwei Seiten Umfang. Wir entschieden uns für einen Umfang von einer Seite pro Stichprobe, da uns die damit erreichte Auswahlmenge von 52 Seiten groß genug für den beabsichtigten Zweck schien und da, sollte sich diese Annahme als falsch erweisen, eine Verdoppelung dieser Menge innerhalb des einmal gewählten Auswahlmodus ohne Schwierigkeiten möglich ist.

Für die beiden ND-Jahrgänge ergeben sich damit folgende Werte[50]:

	ND 1964	ND 1954
Umfang der Zielmenge:	2125 Seiten	1976 Seiten
Aufnahmeintervall:	1 Woche	1 Woche
Zahl der berücks.Ausg.: (= Stichproben)	52	52
Zahl der Seiten: (= 1 Seite p. Stichprobe)	52	52
Auswahlquote:	2,447; abgerundet 2,4%	2,632; abg. 2,6%

In diesen beiden Fällen ist die Aufnahmequote also abhängig vom Aufnahmeintervall und dem Umfang der Stichproben. Dieser Weg mußte eingeschlagen werden, da bisher noch keinerlei Erfahrungswerte für die Ermittlung von Modellmengen dieses Umfangs, insbesondere nicht aus Zeitungstexten, vorlagen.

10.2. Auswahlmodus für WELT 64

Nachdem für die beiden Jahrgänge des ND die Aufnahmequote einmal festlag, mußte bei der Aufnahme des westlichen Vergleichsjahrgangs von dieser, nicht vom Umfang der Stichproben ausgegangen werden; dabei hatte das Aufnahmeintervall gleich zu bleiben. Für WELT 1964 ergibt sich somit folgende Berechnung[51]:

Umfang der Zielmenge	:	5.635 Seiten
Auswahlquote : 2,4%	=	135 Seiten
Aufnahmeintervall	: durchschn.	1 Woche
Zahl der Stichproben	:	52
Umfang der Stichproben	: durchschn.	2,6 Seiten

Da nur ganze Seiten aufgenommen werden, mußte der Umfang der Stichproben entweder 2 oder 3 Seiten betragen[52]. Um die Vergleichbarkeit weiter zu steigern, wurden die gleichen Tagesausgaben wie bei ND 64 zur Aufnahme bestimmt.

11. Modus der Stichprobenverteilung

Die Bestimmung der Ausgaben, aus denen die Stichproben genommen werden, geschah also bei den beiden zuerst aufgenommenen Ost-Jahrgängen nach folgendem Modus:

11.1. Horizontaler Turnus:

Jede Wochentagsausgabe soll berücksichtigt werden; daraus ergibt sich:

1. Woche: Montagsausgabe; 2.Woche: Dienstagsausgabe; 3. Woche: Mittwochsausgabe; usw. (- je 8 Tage Abstand); 6. Woche: Samstagsausgabe; 7. Woche: Montagsausgabe (= 2 Tage Abstand)

11.2. Vertikaler Turnus

Jede S e i t e soll berücksichtigt werden:
1. Ausgabe: Seite 1; 2. Ausgabe: Seite 2; 3. Ausgabe: Seite 3; 4. Ausgabe: Seite 4 usf. bis alle vorhandenen Seiten durchlaufen sind; danach Sprung auf Seite 1.

11.3. Verschiebungen und Abweichungen

a) Verschiebung des Turnus (11.1.) gegenüber Turnus (11.2.):
Unter den gegebenen Bedingungen unseres Materials zeigte es sich, daß beim ND meist eine bestimmte Seite auf die jeweiligen Wochentagsausgaben entfiel, also etwa auf die Montagsausgaben meist die Seite 1, auf die Samstagsausgaben meist die Seite 6 usw. Um dies zu vermeiden, wurde nach Durchlaufen je eines Turnus (also nach 6 Stichproben) der Beginn des neuen Turnus um eine Seite verschoben. Also:
6. Ausgabe: Seite 6: 7. Ausgabe n i c h t Seite 1, sondern Seite 2;
8. Ausgabe S. 3 usf.

b) Ausgleich entsprechend Seitenhäufigkeit:
Gemäß dem unter Punkt 4.5.4.1. dargelegten Verfahren wird für jede Seite ihre Häufigkeit in der Zielmenge absolut und anteilig ermittelt; daraus wird der Anteil jeder Seite an der Auswahlmenge von 52 Seiten (bei ND 64 und ND 54) bzw. 135 Seiten (bei WELT 64) errechnet. Vgl. dazu die Berechnung des Seitenschlüssels im Anhang Ib.

Entsprechend diesen Anteilen wird die nach dem Turnus 11.1, 11.2 und 11.3a ermittelte Verteilung korrigiert[53].

c) Materialbedingte Abweichungen:
Das Archivmaterial, auf das wir uns stützen[54], war leider gelegentlich lückenhaft. So war nicht immer die Berliner Ausgabe des ND vorhanden;

in diesen Fällen wurde auf die Republikausgabe ausgewichen. Fehlte auch diese, wurde die auf die fehlende folgende Ausgabe genommen. Entsprechendes gilt auch für die WELT Berliner Ausgabe. Hier mußte in einigen Fällen die Hamburger Ausgabe genommen werden. Außerdem enthielt der uns übersandte Mikrofilm in 2 Fällen irrtümlich aufgenommene Seiten, nämlich solche aus der Beilage "Reise-Welt". Wir hielten den Fehler jedoch für so geringfügig, daß wir die entsprechenden Seiten in unserem Material belassen haben.

Die nach diesem System aus jedem Jahrgang tatsächlich aufgenommenen Seiten sind im Anhang IIa zusammengestellt.

12. Zusätze zur Modellmenge:

Zum Zwecke statistischer Einzeluntersuchungen (vgl. Teil III) wurde jeweils aus dem Monat Februar jedes der drei Jahrgänge eine bestimmte Zahl von Seiten zusätzlich aufgenommen, um eine höhere Auswahlquote zu erreichen. Aus ND 64-Februar wurden zu den 4 Seiten der Modellmenge weitere 14 Seiten aufgenommen, um eine Auswahlquote von 10% zu erreichen (Gesamtmenge Februar: 184 Seiten)[55]; aus WELT 64-Februar zu den vorhandenen 10 Seiten weitere 12; aus ND 54-Februar zu den vorhandenen 4 Seiten weitere 5.

Ferner wurden zum Zwecke der vergleichenden Untersuchung des Wortschatzes bestimmter prominenter Redner einige Seiten mit offiziellen Reden von W. Ulbricht und einigen anderen Mitgliedern des Politbüros aus ND 64 aufgenommen. Der Modellmenge gehören nur jeweils 1 - 2 Seiten dieser Reden an; sie sind durch den Anhang jetzt jedoch komplett vorhanden.

Schließlich wurden aus verschiedenen Gründen noch einige weitere Seiten aufgenommen, vor allem bei ND 54. Sie sind in einem besonderen Anhang zusammengefaßt.

Umfang und Seitenverteilung der Zusätze siehe Anhang IIb.

13. Besondere Gesichtspunkte bei der Übertragung der Texte auf Datenträger

13.1. Artikel

Die Grundlage der statistischen Textermittlung und -berechnung bildet, wie erwähnt, die Zeitungs s e i t e .

Diese Einheit ist jedoch überwiegend formaler Natur. Darüberhinaus ist das Material in Sinneinheiten gegliedert, die wir "A r t i k e l" nennen, wobei dieser Begriff hier eine erweiterte Bedeutung hat. Unter einem Artikel verstehen wir jede inhaltlich-thematisch zusammengehaltene und (bedingt) formal erkennbare Texteinheit, also auch etwa Wetterbericht, Börsentabellen, Kleinanzeigen, die unter einer Überschrift zusammengefaßt sind, Leserbriefe zum gleichen Thema, Romanfortsetzungen usw. Im einzelnen sind die Grenzen fließend. Sämtliche Artikel, die in einer Modellmenge enthalten sind, erhalten eine fortlaufende Numerierung[56]. Die maschinelle Verarbeitung des Materials orientiert sich vor allem an diesen Artikelnummern, nicht an den Seiten.

13.2. Schreibkonventionen[57]

Abgeschrieben, d.h. auf Lochkarten übertragen[58] werden alle auf den Seiten vorhandenen Texte mit Ausnahme

a) des Zeitungskopfes
b) des Impressums
c) reiner Zahlentabellen
d) von Bildern, Zeichnungen usw.

In den Fällen c) und d) wird ein Vermerk aufgenommen, was an diesen Stellen übergangen wurde[59].

13.2.1. Informationskonstanz bei einzelnen Wörtern und Zeichen

Texte bestehen aus Buchstabenfolgen, Zeichen (Satzzeichen) und Zahlen (die letzteren sind hier irrelevant) sowie aus Kombinationen dieser Gattungen.
Zahlreiche Buchstabenfolgen ("Wörter") und Zeichen sind nur im Textzusammenhang eindeutig, isoliert jedoch mehrdeutig.

Da unsere Datenverarbeitung zunächst auf die Isolierung der einzelnen "Wörter" und Zeichen abzielt, wird mittels bestimmter Vorschriften versucht, einen Teil der entstehenden Informationsverluste aufzufangen, soweit sie für die maschinelle Verarbeitung von Bedeutung sind.

Satzzeichen und Sonderzeichen:

Da eine Reihe von Zeichen auf den Lochkartenschreibern fehlen, müssen Ersatzzeichen[60] verwendet werden:

Doppelpunkt	= / 0
Fragezeichen	= / 1
Anführungszeichen	= / 2
Ausrufezeichen	= / 3
Semikolon	= / 4
einf. Anführungszeichen	= / 5
Prozentzeichen	= / 6
Paragraphenzeichen	= / 7
öffnende eckige Klammer	= / 8
schließende eckige Klammer	= / 9
Abkürzungspunkt	= '.
Auslassungspunkte	= ' ' '

Im übrigen gelten die schon in den "Schreibanweisungen für die Übertragung von Texten auf Lochstreifen" - Mannheimer Bericht - unter Punkt 5a, 7, 9 - 11, 13 - 17 erwähnten Vorschriften.

Abweichend gilt[61]: Großschreibung (nur bei Substantiven und großgeschriebenen Anreden) wird durch das Zeichen * (Sternchen) vor dem Wort markiert; großgeschriebene Adjektive (vor allem als Namensbestandteile) werden durch '* (Apostroph-Stern) vor dem Wort gekennzeichnet; Personen-Familiennamen erhalten zu dem Großschreibungszeichen noch das Kennzeichen Stern-Stern-blank vor dem Wort.

Unberücksichtigt bleiben semantische oder inhaltliche sowie grammatische Mehrdeutigkeiten.
Ebenfalls unberücksichtigt bleiben Schriftartunterschiede (Kursiv-, Sperr-, Fett- oder Petitdruck usw.), diakritische Zeichen, der Unterschied zwischen römischen und arabischen Zahlen (römische werden in arabische überführt[62] sowie der Unterschied zwischen "Anführungsstriche unten" und "Anführungsstriche oben".

13.2.2. Informationskonstanz bei Wortgruppen

Bei der maschinellen Zerlegung unserer Texte in einzelne Wörter und Zeichen treten weitgehende Informationsverluste auf, die teils die Bedeutung bestimmter Gruppen und Wörter, teils charakteristische Eigenheiten größerer Textteile betreffen.
Diese Informationsverluste werden zum Teil durch ein System von Transkriptionszeichen aufgefangen[63]. Es entspricht im wesentlichen dem in Mannheim verwendeten System (vgl. "Schreibanweisungen" Punkt 6a-f)[64].

Zusätzlich werden transkribiert: Personeneigennamen samt ihren Titeln und Anreden, Sach-Eigennamen (nur mehrgliedrige)[65], übersetzte Textstellen, Abkürzungen (soweit nicht mit Punkt abgekürzt)[66] sowie Textstellen, die nicht schreibbar sind bzw. wegen Irrelevanz (Zahlenkolonnen) nicht geschrieben werden sollen[67].
Die durch die Transkription dem Text beigefügten Informationen bleiben bei der Zerlegung den einzelnen Wörtern erhalten.

Ebenso können die zu einer transkribierten Textstelle gehörigen Wörter wieder zusammengefügt werden.

13.2.3. Informationskonstanz des Artikel-Charakters

Der Zeitungsleser entnimmt jedem Zeitungsartikel als ganzem schon bei flüchtigem Anschauen eine Reihe von Informationen, die ebenfalls bei der Zerlegung verlorengehen.
Ein Teil dieser Informationen, die sich auf den ganzen Artikel beziehen, wird auf "Informationskarten" gespeichert, d.h. chiffrierten Lochkarten, die jedem Artikel vorangestellt werden[68]. Die Informationskarte (IK) enthält (außer einer Chiffre für die Kartenart und der Nummer des Artikels, auf den sich die IK bezieht) folgende Angaben:

1. Chiffre der Zeitung mit Jahrgang, Monat, Tag, Seite
2. Herkunft der Zeitung (West oder Ost)
3. Aufnahmeprinzip (Statistisches = Modellmenge)
4. Art der Zeitung (Tages-, Wochen- usw.)
5. Angabe, ob es sich um ein Artikelfragment handelt[69]
6. Verfasser des Artikels (falls angegeben)
7. Agenturen (bis zu drei)
8. Sachgebiete (bis zu drei gleichzeitig)[70]
9. Innere Form
   a) Publizistisches Ziel (bis zu zwei gleichzeitig)[71]
   b) Artikelart bzw. -gattung mit Sonderformen
   c) Sparte, der der Artikel eingeordnet ist
10. Äußere Form (Aufmachung)
    a) Breite (in Spalten)
    b) Länge (in Zeilen)
11. Angaben über stilistische Kriterien (nicht ausgefüllt).

Auf einer weiteren Karte, der "Beikarte" (BK) oder "Themakarte" wird, außer der Artikelnummer, das Thema, das in dem Artikel behandelt wird, angegeben. Die Karte ist gegliedert in einen "Kopfteil", der das "Land" oder die "Länder" angibt, auf das sich das behandelte Thema bezieht, und einen "Textteil", der das Thema in Stichworten wiedergibt.

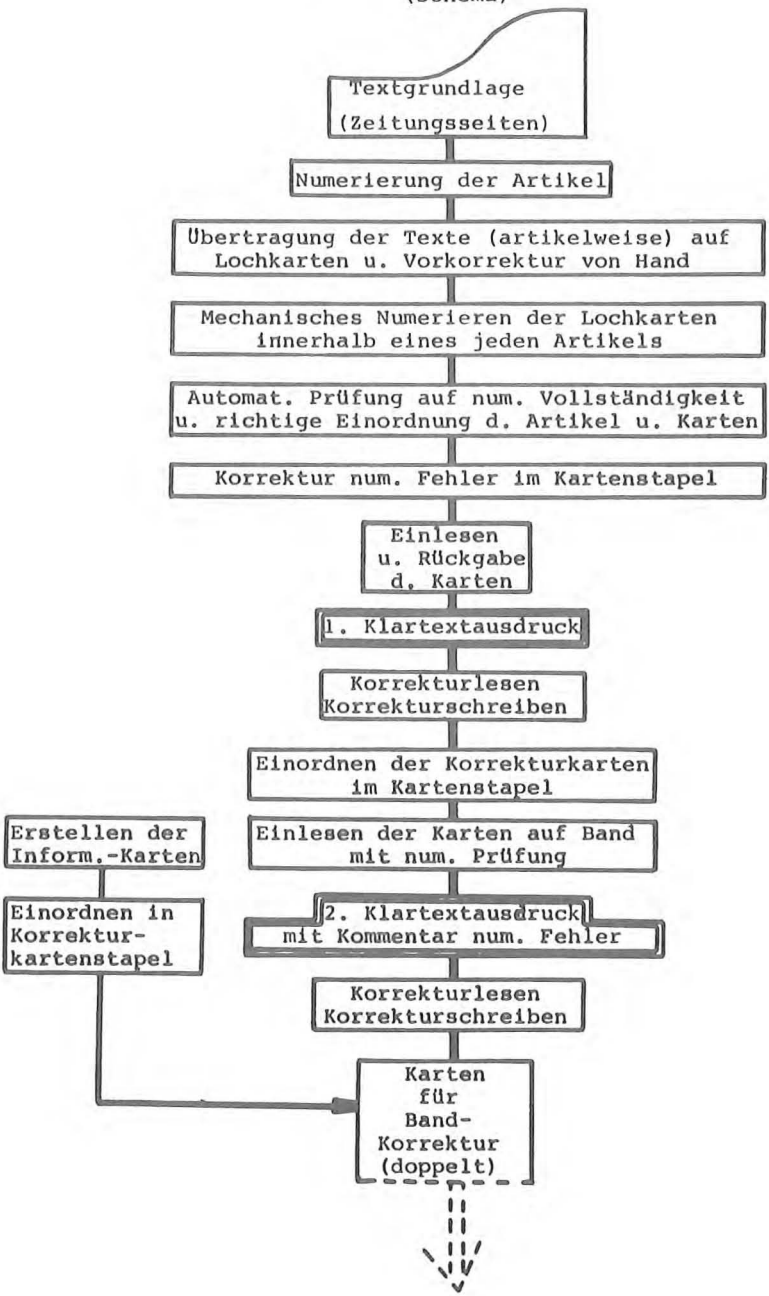
Die Angaben dieser beiden Karten werden in nicht chiffrierter, dafür begrenzter Form zusätzlich auf einer Handkartei (Artikelkartei) festgehalten.

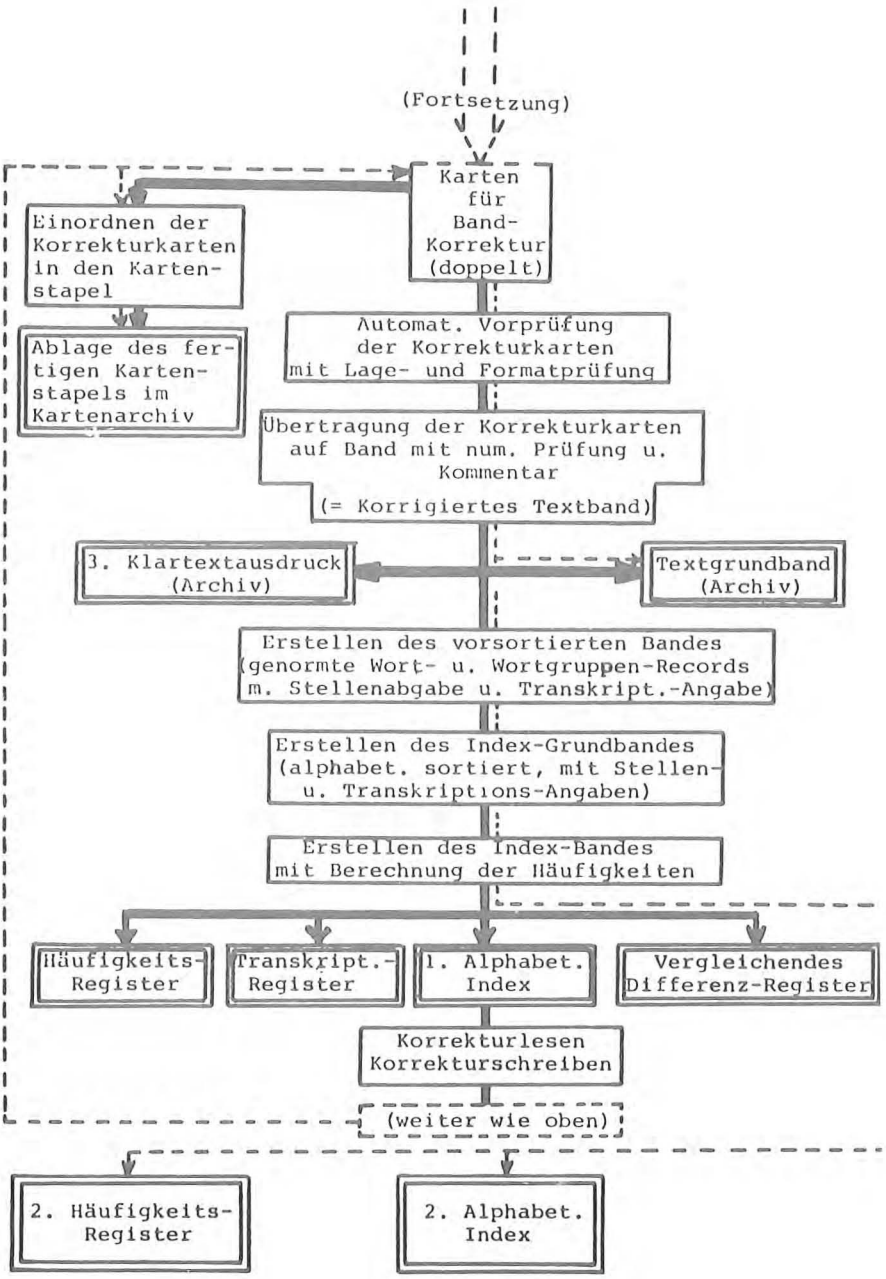
Auf einer dritten Lochkarte ("Zusatz"- oder "Verweiskarte") (ZK) können Artikel angegeben werden, die mit dem behandelten inhaltlich oder thematisch korrespondieren oder ihn ergänzen (falls es sich um ein Fragment gehandelt hat) [72].

Eine Übersicht über den Aufbau der Informationskarte siehe im "Auszug aus den Erläuterungen zur Informationskarte" im Anhang III.

Arbeitsgang

(Schema)




(Fortsetzung)
Karten für Band-Korrektur (doppelt)
Einordnen der Korrekturkarten in den Kartenstapel
Ablage des fertigen Kartenstapels im Kartenarchiv
Automat. Vorprüfung der Korrekturkarten mit Lage- und Formatprüfung
Übertragung der Korrekturkarten auf Band mit num. Prüfung u. Kommentar
(= Korrigiertes Textband)
3. Klartextausdruck (Archiv)
Textgrundband (Archiv)
Erstellen des vorsortierten Bandes (genormte Wort- u. Wortgruppen-Records m. Stellenabgabe u. Transkript.-Angabe)
Erstellen des Index-Grundbandes (alphabet. sortiert, mit Stellen- u. Transkriptions-Angaben)
Erstellen des Index-Bandes mit Berechnung der Häufigkeiten
Häufigkeits-Register
Transkript.-Register
1. Alphabet. Index
Vergleichendes Differenz-Register
Korrekturlesen Korrekturschreiben
(weiter wie oben)
2. Häufigkeits-Register
2. Alphabet. Index

14.1. Erläuterungen und technische Angaben
zum Arbeitsablauf und zu den verwendeten Geräten

1. Ablochen. Wir benutzen die IBM-80-Spalten-Standard-Lochkarte in zwei Farben (für Ost- und Westtexte)[73].

   Aufbau einer Textkarte: Sp. 1-4 = Artikelnummer
   Sp. 5-7 = Kartennummer
   Sp. 8 = blank
   Sp. 9-80 = fortlaufender Text, ohne Rücksicht auf Zeitungs-Zeilenlänge.

   Geräte[74]: zwei IBM-Schreiblocher 026 (mit Textzeile am oberen Rand der Karten)[75].

   Das Gerät verfügt über 46 Zeichen (Symbolausführung H).

2. Mechanisches Numerieren. Die Artikelnummer wird schon beim Ablochen in die Karten dupliziert, die laufende Kartennummer im Doppler IBM Typ 519 artikelweise hinzugefügt (gestanzt und aufgedruckt).

3. Automatische numerische Prüfung. Programm PRUEF des IDS mit Unterprogrammen (vgl. Anhang IV, Programmübersicht Nr. 1).

4. Klartextausdrucke. Der Klartextausdruck gibt den Text kartenweise wieder (jede Ausdruckzeile entspricht einer Karte). Jeder Artikel beginnt auf einer neuen Seite. Unregelmäßigkeiten in Numerierung und Kartenformat werden kommentiert. Der Klartextausdruck kann eine für das Korrigieren eingerichtete Form haben.
   Am Ende eines jeden Artikels erscheint ein Schreibervermerk (Namenssigle und Datum des Schreibens).

5. Erste Korrekturphase. Die Texte werden nur durch Austausch der falschen gegen die neuen richtigen Karten im Kartenstapel korrigiert. Erst danach werden die Texte insgesamt auf Band gelesen und dabei nochmals numerisch geprüft.

6. Zweite Korrekturphase (auf Magnetband).

Vorprüfung: Programm VORKOR (Programmübersicht Nr. 2)
Korrektur: Programm KORR und NDNUM (Programmübersicht Nr. 3 und 4)

In der zweiten Korrekturphase werden meist auch die Informationskarten (sowie Bei- und Zusatzkarten) hinzugefügt (jeweils als erste Karte jedes Artikels).

7. Textband

Band: IBM MC1.800, Länge 1000 feet
Schreibdichte: 556 b.p.i.
Recordlänge: 14 á 6
Blockung: variabel, gewöhnlich 1400
Schreibmodus: wählbar, gewöhnlich BCD
Allgemein: Siehe Programmbeschreibung EINAUS des IPK Bonn (Jahresbericht des IPK 1966/7, S. 39)

Dieses Textband ist Datenträger für die Suchwort-Programmläufe.

8. Erstellen des Index. Siehe "Programmübersicht" Nr. 16-20

8a. Der alphabetische Wortformen-Index enthält alle Wortformen mit Stellenangaben (Artikel- und Zeilennummer), ferner Angaben zur etwaigen Transkription des einzelnen Wortes zu jeder Fundstelle sowie relative und absolute Häufigkeit. Bestimmte oder alle Transkriptionen können dabei ausgefiltert werden. Bindestrich-Komposita werden unter allen ihren Teilwörtern aufgeführt, jedoch nur einmal gezählt.

8b. Index nach Kriterien der Informationskarten. Es können Indices aus Artikeln erstellt werden, auf die ein oder mehrere gewünschte Kriterien der Informationskarte zutreffen. Gleiches gilt für Indices nach Beikarten (Indices zu Artikeln über bestimmte Themen). (Programmübersicht Nr. 5)

9. Transkriptionsregister. Vom Index-Grundband ausgehend können alphabetische Spezial-Verzeichnisse zu allen Transkriptionen in Form alphabetischer Wortformen-Indices oder in Form von Wortgruppen-Registern erstellt werden (Programmübersicht Nr. 19).

10. Kombinierbarkeit. Die Auswahl- bzw. Spezialisierungsverfahren nach Kriterien der Informationskarten und nach Transkriptionen können kombiniert werden[76]. Die Zahl der sich aus der Kombination der vielfachen Abfragemöglichkeiten der Informations- und Beikarte mit den Möglichkeiten des Transkriptionssystems ergebenden sinnvollen und realisierbaren speziellen Verzeichnisse liegt bei mehreren tausend (die Zahl der theoretischen Möglichkeiten ist nicht genau angebbar, da einige Chiffrengruppen laufend erweiterbar sind; sie liegt aber über 2 Milliarden).

11. Vergleichende Differenzregister zur Ermittlung des abweichenden oder gemeinsamen Wortschatzes zweier Texte liegen noch nicht vor; das Programm ist in Arbeit. (Programm-Übersicht Nr. 21)

12. Dritte Korrekturphase. Erfahrungsgemäß zeigen sich manche Schreibfehler erst im Index[77]. Nochmalige Korrektur über alle Stufen der Verarbeitung hinweg ist also notwendig. Nach diesem Korrekturgang liegt die Fehlerquote in unseren Texten[78] unter 0,5 Promille (wahrscheinlich bei 0,2 - 0,1 Promille).
Dies dürfte ein bei großen Textmengen in der Praxis äußerst hoher Grad von Fehlerfreiheit sind. Er liegt selbstverständlich weit über der Fehlerfreiheit der Zeitungstexte selbst (Druckfehler im Zeitungstext werden beim Ablochen berichtigt).

15. Übersicht über den Stand der Arbeiten

(Stand von Frühjahr 1968)

15.1. Texte

1. Neues Deutschland 1964
   a) Modellmenge:
      52 Seiten, 680 Artikel, ca. 180.000 lfd. Wörter
      Dieser Text lag einschließlich Informationskarten während des Jahres 1967 fertig zur Benutzung vor; es bestehen ein alphabetischer Index sowie Teilindices, ein Häufigkeitsregister und einige Transkriptionsregister.
   b) Zusatzmenge 1 ("Februarmenge")
      14 Seiten, 140 Artikel ca. 50.000 lfd. Wörter
      Benutzbar seit Sommer 1967, Informationskarten teilweise ausgefüllt.
   c) Zusatzmenge 2 (offizielle Reden)
      13 Seiten, 25 Artikel, ca. 46.000 lfd. Wörter
      Benutzbar seit Frühjahr 67. Informationskarten teilweise ausgefüllt.
   d) Zusatzmenge 3 (sonstige Seiten)
      3 Seiten, 30 Artikel, ca. 10.000 lfd. Wörter
      Benutzbar seit Sommer 67. Informationskarten fertig ausgefüllt.

   Das Gesamtkorpus aus ND 64 umfaßt somit 82 Seiten mit über 280.000 lfd. Wörtern.
   Ein Gesamtindex und Häufigkeitsregister zu allen vier Einzelmengen gemeinsam werden vorbereitet.

2. DIE WELT 1964
   a) Modellmenge:
      135 Seiten, 1640 Artikel, ca. 410.000 lfd Wörter
      Der Text liegt seit Ende 1967 korrigiert zur Benutzung vor. Informationskarten sind zu ca. 1/3 erstellt. Es sind ein Teilindex sowie Teilregister von Transkriptionen vorhanden.

b) Zusatzmenge 1 ("Februarmenge")
12 Seiten, ca. 125 Artikel, ca. 35.000 lfd. Wörter
Korrigiert benutzbar seit Anfang 1968 (noch ohne Informationskarten).

c) Zusatzmenge 2 (offizielle Reden)
entfällt.

d) Zusatzmenge 3 (sonstige Seiten)
5 Seiten, 40 Artikel, ca. 14.000 lfd. Wörter
Korrigiert benutzbar seit Frühjahr 68 (noch ohne Informationskarten).

Das Gesamtkorpus umfaßt somit 152 Seiten mit ca. 460.000 lfd. Wörtern.

3. Neues Deutschland 1954
Dieser Text wurde noch vor dem Jahrgang WELT 64 auf Lochkarten übertragen, jedoch erst nach der WELT 64 korrigiert.

a) Modellmenge
52 Seiten, 610 Artikel, ca. 165.000 lfd. Wörter
Der Text lag im Frühjahr 1968 einmal korrigiert und somit begrenzt zur Benutzung vor (noch ohne Informationskarten). Indices sind im Herbst 1968 zu erwarten.

b) Zusatzmenge 1 ("Februarmenge")
5 Seiten, 62 Artikel, ca. 16.000 lfd. Wörter
Benutzbarkeit wie unter a)

c) Zusatzmenge 2 (offizielle Reden)
entfällt

d) Zusatzmenge 3 (sonstige Seiten)
11 Seiten, ca. 130 Artikel, ca. 35.000 lfd. Wörter
Benutzbarkeit wie unter a).

Das Gesamtkorpus aus ND 54 umfaßt also 68 Seiten mit ca. 216.000 Wörtern. Es wird voraussichtlich im Herbst 1968 einschließlich Informationskarten voll zur Verfügung stehen.

4. DIE WELT 1954

Mit der Aufnahme der Texte aus WELT 54 wird im Spätsommer 1968 begonnen werden. Die Auswahl der Texte wird voraussichtlich 90 Seiten umfassen.

Erweiterungen

Es ist zu erwarten, daß in absehbarer Zeit westliche Zeitungstexte aus der laufenden Produktion über Zeitungslochstreifen (TTS-Streifen) unmittelbar auf Band genommen werden können. Ein Umwandlungsprogramm, das die Streifentexte in eine unseren Schreibkonventionen stark angenäherte Form bringt, ist in der Außenstelle erstellt worden[79].

15.2. Auswertung

Die Außenstelle hat sich, ihrer Aufgabe entsprechend, auf solche Auswertungsprogramme konzentriert, die für Wortschatzuntersuchungen erforderlich sind. Um die Bonner Texte jedoch auch für Mannheimer syntaktische Untersuchungen und umgekehrt verwertbar zu machen, ist in der Außenstelle ein Umwandlungsprogramm erstellt worden, das die Bonner Textbänder auf die für die Mannheimer Bänder vom Rechenzentrum Darmstadt geforderten Konventionen umstellt bzw. Mannheimer Bänder auf die in Bonn gültigen Konventionen umstellt (vgl. Programm-Übersicht Nr. 12 und 13).

Zur Auswertung der Texte in Form von Indices oder Transkriptionsregistern, die mit Abfragen nach Kriterien der Informationskarte gekoppelt sein können, vgl. Punkt 14.1., Abschnitt 8-10. Solche Register aufgrund kombinierter Abfragen wurden auf den Tagungen des Instituts im Herbst 1965 und Frühjahr 1968 vorgelegt (vgl. Anmerkung 76).

Als wesentlich wichtiger für Arbeiten zum Wortschatz hat sich das unter Nr. 6 und 7 der Programm-Übersicht kurz beschriebene System "MOSU" zur Suche von Wörtern mit Kontext erwiesen. Es können aufgrund vorzugebender Buchstabenkombinationen ganze Wörter oder Wortteile, d.h. Wort"stämme", -anfänge, -endungen, gesucht werden [80]. Einschränkung auf Suche nach Substantiven oder Nicht-Substantiven ist möglich. Alle gefundenen Wörter werden gesondert sowie in ihrem Kontext (variabel zwischen 1 und 9 Zeilen Kontext oder 2 - 7 Wörter Kontext oder 1 Satz Kontext) ausgedruckt. Jedem Beleg können Angaben aus der Informationskarte (dechiffriert) beigegeben werden [81]. Der Kontext kann von allen störenden Zusatz- oder Transkriptionszeichen gereinigt werden.

Mit Programmläufen dieser Art hat die Außenstelle in den Jahren 1966 und 1967 18 verschiedene wissenschaftliche Arbeitsvorhaben unterstützt. Der Umfang der ausgegebenen Menge beträgt über 300.000 Zeilen Kontext. Unter den 12 im Jahre 1967 geförderten Arbeitsvorhaben befanden sich 4 von ehemaligen oder jetzigen Mitarbeitern des Instituts in Mannheim, Bonn und Innsbruck, 4 betrafen Staatsexamensarbeiten oder Dissertationen der Bonner Universität, 4 weitere betrafen Arbeitsvorhaben von anderen Wissenschaftlern. Ein Teil der Arbeiten ist abgeschlossen.

Das Programmsystem wird weiter ausgebaut und verfeinert werden.

Anhang I a / 1

Seitenberechnung für ND 1964:

Ausgegangen wurde, wie unter Punkt 10.1. beschrieben, vom Umfang der Auswahlmenge, d.h. von einer Seite pro Zeitungswoche = 52 Seiten.
Ausgezählt wurden 153 Erscheinungsmengen aus 5 Monaten (Januar, Februar, April, Juli, Oktober) mit 1200 Seiten, bzw. 132 Ausgaben (d.h. ohne Sonntagsausgaben und Beilagen) mit 912 Seiten.
Auf 360 Jahres-Erscheinungsmengen umgerechnet ergibt sich daraus eine Gesamt-Bezugsmenge S_{eJ} von 2820 Seiten; auf 308 Jahrgangs-Ausgaben umgerechnet eine Zielmenge S_{aJ} von 2125 (vgl. Mengenbeschreibung unter Punkt 8.2.) 8
52 Seiten von 2125 sind 2,45 %, $\sim$ 2,4 %.

Aufstellung der ausgezählten Ausgaben verschiedener Länge in Klassen:

Klasse	Länge d. Ausgaben in Seiten (y bis y+m)	Zahl der Ausgaben (p bis p_i)	Zahl der Seiten pro Klasse (p mal y)	Häufigkeit (F) jeder Seite in 912 S. ausgez. Menge
1	4	17	68	4 x 132
2	5/6	46	276	2 x 115
3	7/8	65	520	2 x 69
4	9/10	--	--	2 x 4
5	11/12	4	48	2 x 4
		n=132	912	

Entsprechend der unter 4.5.4.2. ermittelten Formel $f = F.\frac{S_M}{S_{aJ}}$ wird die Anzahl jeder Seite pro Klasse in der Auswahlmenge von 52 Seiten ermittelt; dabei ist $\frac{S_M}{S_{aJ}}$ hier = 0,057 [83].

Anhang 1 a / 2

Seitenberechnung für WELT 1964

Ausgezählt wurden 131 Ausgaben aus 5 Monaten (Januar, Februar, April, Juli, Oktober) mit insgesamt 2429 Seiten[84]. Auf das Jahr, d.h. auf 304 Ausgaben bezogen, ergibt sich eine Zielmenge von ∼ 5635 Seiten (vgl. Mengenbeschreibung Punkt 8.2.).

Bei einer Auswahlquote von 2,4 % muß der Umfang der Auswahlmenge somit 135 Seiten betragen.

Aufstellung der ausgezählten Ausgaben verschiedener Länge in Klassen:

Klasse	Länge d. Ausgaben in Seiten (y bis y+m)	Zahl der Ausgaben (p bis p_i)	Zahl d. Seiten pro Klasse (p mal y)	Häufigkeit (F) jeder Seite in 2428 ausgez. Menge
1	12	10	120	12 x 131
2	13/14	26	364	2 x 121
3	15/16	35	560	2 x 95
4	17/18	21	378	2 x 60
5	19/20	8	160	2 x 39
6	21/22	10	220	2 x 31
7	23/24	0	0	2 x 21
8	25/26	7	182	2 x 14
9	27/28	4	112	2 x 14
10	29/30	1	30	2 x 10
11	31/32	3	96	2 x 9
12	33/34	5	170	2 x 6
13	35/36	1	36	2 x 1
		n=131	2428	

Entsprechend der Formel $f = F \cdot \frac{S_M}{S_{aJ}}$

wird die Anzahl jeder Seite pro Klasse in der Auswahlmenge von 135 Seiten ermittelt; dabei ist $\frac{S_M}{S_{aJ}}$ hier = 0,0556.

Anhang I a / 3

Seitenberechnung für ND 1954:

Ausgegangen wurde, wie bei ND 64, vom Umfang der Auswahlmenge, d.h. von einer Seite pro Zeitungswoche = 52 Seiten.

Ausgezählt wurden 104 Erscheinungsmengen aus 4 Monaten (Januar, April, Juli, Oktober) mit 740 Seiten bzw. ebenfalls 104 Ausgaben[85)] (ohne Beilagen) mit 676 Seiten.

Auf 304 Jahreserscheinungsmengen umgerechnet ergibt sich daraus eine Gesamtmenge S_{eJ} von 2163 Seiten; auf 304 Jahrgangsausgaben umgerechnet eine Zielmenge S_{aJ} von 1976 Seiten (vgl. Mengenbeschreibung, Punkt 8.2.). 52 Seiten von 1976 Seiten sind 2,63 %.

Aufstellung der ausgezählten Ausgaben verschiedener Länge in Klassen:

Klasse	Länge d. Ausgaben in Seiten (y bis y+m)	Zahl der Ausgaben (p bis p_i)	Zahl d. Seiten pro Klasse (p mal y)	Häufigkeit (F) jeder Seite in 676 S. ausgezählter Menge
1	4	10	40	4 x 104
2	5/6	59	354	2 x 94
3	7/8	34	272	2 x 35
4	9/10	1	10	2 x 1
		n = 104	676	

Entsprechend der Formel $f = F \cdot \frac{S_M}{S_{aJ}}$ wird die Anzahl jeder Seite pro Klasse in der Auswahlmenge von 52 Seiten ermittelt; dabei ist $\frac{S_M}{S_{aJ}}$ hier = 0.0769.

Anhang I b Berechnung der Häufigkeit jeder Seite in der Auswahlmenge

Seite	Neues Deutschland 64 Vork. in der ausgez. Menge (912 S.)	Vork. in der Auswahlmenge (52 S.) F mal 0,057 = f rechn.	rund	Seite	DIE WELT 64 Vork. in der ausgez. Menge (2429 S.)	Vork. in der Auswahlmenge (135 S.) F mal 0,0556 = f rechn.	rund
1	132	7,52	7	1	131	7,28	8
2	132	7,52	7	2	131	7,28	7
3	132	7,52	7	3	131	7,28	7
4	132	7,52	7	4	131	7,28	7
5	115	6,56	7	5	131	7,28	8
6	115	6,56	7	6	131	7,28	7
7	69	4,00	5	7	131	7,28	7
8	69	4,00	4	8	131	7,28	7
9	4	0,23	-	9	131	7,28	8
10	4	0,23	1	10	131	7,28	7
11	4	0,23	-	11	131	7,28	7
12	4	0,23	-	12	131	7,28	7
				13	121	6,73	7
	912		52	14	121	6,73	7
				15	95	5,28	5
				16	95	5,28	5
				17	60	3,34	4
				18	60	3,34	3
				19	39	2,17	2
				20	39	2,17	2
				21	31	1,72	2
				22	31	1,72	2
				23	21	1,17	1
				24	21	1,17	1
				25	21	1,17	1
				26	21	1,17	1
				27	14	0,78	1
				28	14	0,78	1
				29	10	0,56	1
				30	10	0,56	-
				31	9	0,50	1
				32	9	0,50	1
				33	6	0,33	-
				34	6	0,33	-
				35	1	0,05	-
				36	1	0,05	-
					2.428		135

Seite	Neues Deutschland 54 (676 S.) Vork. in der ausgez. Menge	(52 S.) Vork. in der Auswahlmenge F mal 0,0769 = f rechn.	rund
1	104	8,00	8
2	104	8,00	8
3	104	8,00	8
4	104	8,00	8
5	94	7,23	8
6	94	7,23	7
7	35	2,69	3
8	35	2,69	3
9	1	0,08	-
10	1	0,08	-
	676		52

Liste der für die Modellmengen aufgenommenen Seiten

	N D 1964		DIE W E L T 1964		N D 1954		DIE W E L T 1954	
Nr.	Datum	Seite	Datum	Seiten	Datum	Seite	Datum	Seiten
1)	Do 2. 1.	1	Do 2. 1.	1, 2, 4	Fr 1. 1.	1	Fr 1. 1.	
2)	Fr 10. 1.	2	Fr 10. 1.	3, 4	So 10. 1.	3	Mo 11. 1.	
3)	Sa 18. 1.	3	Sa 18. 1.	5, 6	Di 19. 1.	4	Di 19. 1.	
4)	Mo 20. 1.	2	Mo 20. 1.	7, 8, 9	Mi 27. 1.	5	Mi 27. 1.	
5)	Di 28. 1.	3	Di 28. 1.	10,11,12	Do 4. 2.	7	Do 4. 2.	
6)	Mi 5. 2.	4	Mi 5. 2.	13,14	Fr 12. 2.	2	Fr 12. 2.	
7)	Do 13. 2.	5	Do 13. 2.	15,16	Sa 20. 2.	3	Sa 20. 2.	
8)	Fr 21. 2.	6	Fr 21. 2.	1, 2, 3	So 21. 2.	4	Mo 22. 2.	
9)	Sa 29. 2.	7	Sa 29. 2.	3, 4,11	Di 2. 3.	5	Di 2. 3.	
10)	Mo 2. 3.	3	Mo 2. 3.	1, 5, 6	Mi 10. 3.	6	Mi 10. 3.	
11)	Di 10. 3.	4	Di 10. 3.	7, 8,12	Do 18. 3.	1	Do 18. 3.	
12)	Mi 18. 3.	5	Mi 18. 3.	9,10	Fr 26. 3.	3	Fr 26. 3.	
13)	Do 26. 3.	6	Do 26. 3.	11,12,13	Sa 3. 4.	4	Sa 3. 4.	
14)	Fr 3. 4.	7	Fr 3. 4.	15,16	So 4. 4.	7	Mo 5. 4.	
15)	Sa 11. 4.	5	Sa 11. 4.	17,18,31,32	Di 13. 4.	5	Di 13. 4.	
16)	Mo 13. 4.	4	Mo 13. 4.	1,13,14	Mi 21. 4.	6	Mi 21. 4.	
17)	Di 21. 4.	5	Di 21. 4.	19,20	Do 29. 4.	6	Do 29. 4.	
18)	Mi 29. 4.	6	Mi 29. 4.	21,22	Fr 7. 5.	4	Fr 7. 5.	
19)	Fr 15. 5.	1	Fr 15. 5.	3, 4,13	Sa 15. 5.	5	Sa 15. 5.	
20)	Sa 23. 5.	2	Sa 23. 5.	5, 6,24,28	So 16. 5.	2	Mo 17. 5.	
21)	Mo 25. 5.	3	Mo 25. 5.	7, 8	Di 25. 5.	1	Di 25. 5.	
22)	Di 2. 6.	6	Di 2. 6.	5, 9,10	Mi 2. 6.	2	Mi 2. 6.	
23)	Mi 10. 6.	7	Mi 10. 6.	11,12	Do 10. 6.	3	Do 10. 6.	
24)	Do 18. 6.	8	Do 18. 6.	2,13,14	Fr 18. 6.	5	Fr 18. 6.	
25)	Fr 26. 6.	1	Fr 26. 6.	1,15,16	Sa 26. 6.	8	Sa 26. 6.	
26)	Sa 4. 7.	2	Sa 4. 7.	17,18,26,27	So 27. 6.	7	Mo 28. 6.	

noch nicht ermittelt

Nr.	N D 1964 Datum	Seite	DIE W E L T 1964 Datum	Seiten	N D 1954 Datum	Seite	DIE W E L T 1954 Datum	Seiten
27)	Mo 6. 7.	4	Mo 6. 7.	3, 4	Di 6. 7.	6	Di 6. 7.	
28)	Di 14. 7.	1	Di 14. 7.	5, 6	Mi 14. 7.	1	Mi 14. 7.	
29)	Mi 22. 7.	2	Mi 22. 7.	6, 7, 8	Do 22. 7.	2	Do 22. 7.	
30)	Do 30. 7.	3	Do 30. 7.	9,10	Fr 30. 7.	6	Fr 30. 7.	
31)	Fr 7. 8.	8	Fr 7. 8.	5,13,14	Sa 7. 8.	1	Sa 7. 8.	
32)	Sa 15. 8.	6	Sa 15. 8.	21,22	So 8. 8.	2	Mo 9. 8.	
33)	Mo 17. 8.	1	Mo 17. 8.	11,12	Di 17. 8.	3	Di 17. 8.	
34)	Di 25. 8.	2	Di 25. 8.	1, 2, 8	Mi 25. 8.	4	Mi 25. 8.	
35)	Mi 2. 9.	3	Mi 2. 9.	1, 2	Do 2. 9.	5	Do 2. 9.	
36)	Do 10. 9.	4	Do 10. 9.	3, 4, 7	Fr 10. 9.	1	Fr 10. 9.	
37)	Fr 18. 9.	5	Fr 18. 9.	5, 6	Sa 18. 9.	2	Sa 18. 9.	
38)	Sa 26. 9.	8	Sa 26. 9.	23,25	So 19. 9.	3	Mo 20. 9.	
39)	Mo 28. 9.	6	Mo 28. 9.	7, 8	Di 28. 9.	4	Di 28. 9.	
40)	Di 6.10.	10	Di 6.10.	9,10	Mi 6.10.	5	Mi 6.10.	
41)	Mi 14.10.	4	Mi 14.10.	2,11,17	Do 14.10.	6	Do 14.10.	
42)	Do 22.10.	2	Do 22.10.	13,14,17	Fr 22.10.	1	Fr 22.10.	
43)	Fr 30.10.	1	Fr 30.10.	14,15,16	Sa 30.10.	2	Sa 30.10.	
44)	Sa 7.11.	7	Sa 7.11.	9,19,20	So 31.10.	4	Mo 1.11.	
45)	Mo 9.11.	3	Mo 9.11.	1, 2,10	Di 9.11.	8	Di 9.11.	
46)	Di 17.11.	4	Di 17.11.	3, 4	Mi 17.11.	3	Mi 17.11.	
47)	Mi 25.11.	5	Mi 25.11.	5, 6, 9	Do 25.11.	4	Do 25.11.	
48)	Do 3.12.	8	Do 3.12.	16,17	Fr 3.12.	5	Fr 3.12.	
49)	Fr 11.12.	7	Fr 11.12.	7, 8	Sa 11.12.	6	Sa 11.12.	
50)	Sa 19.12.	6	Sa 19.12.	14,19,16,25	So 12.12.	1	Mo 13.12.	
51)	Mo 21.12.	1	Mo 21.12.	9,10	Di 21.12.	2	Di 21.12.	
52)	Di 29.12.	5	Di 29.12.	11,12	Mi 29.12.	3	Mi 29.12.	

noch nicht ermittelt

ANHANG II b

Liste der zusätzlich aufgenommenen Seiten

a) Seiten zur Verdichtung der jeweiligen Februarmenge

N D 64		DIE W E L T 64		N D 54		DIE W E L T 54	
Datum	Seiten	Datum	Seiten	Datum	Seite	Datum	Seiten noch nicht festgelegt
Fr 31. 1.	2,3	Sa 1.2.	1,2,4	So 31. 1.	1	Mo 1.2.	
Di 5. 2.	1	Mo 10.2.	5,6,7	Di 9. 2.	5	Mo 8.2.	
Mo 10. 2.	2,3	Mo 17.2.	8,9,10	Di 16. 2.	6	Di 16.2.	
Do 13. 2.	6,8	Di 25.2.	11,12,13	Mi 24. 2.	8	Fr 26.2.	
Mo 17. 2.	2,7			So 28. 2.	4		
Fr 21. 2.	4						
Di 25. 2.	4,5						
Sa 29. 2.	1,8						

b) Seiten zur Ergänzung vorhandener Artikel (offizielle Reden)

Di 5. 2.	(1),3 5,6,7		
Do 13. 2.	3,4		
Do 6. 2.	3,4,5		
Sa 29. 2.	(1),4, 5,6,8		
So 1. 3.	5,6	(Einklammerung = teilweise)	

c) Sonstige zusätzliche Seiten

Di 22. 1. 2	Mi 22.1. 1,2	So 11. 4. 7	
Di 2. 9. 8	Do/Fr 7./8.5. 1,2	Sa 9. 1. 8	Sa 9.1. .....
Di 6.10. 3	Sa 11.4. 2	So 17. 1. 3	
		Do 29. 4. 8	
		So 23. 5. 6	
		So 4. 7. 6,7	
		So 15. 8. 2	
		So 26. 9. 3	
		So 7.11. 7	
		So 19.12. 1	

Anhang III

Auszug aus den Erläuterungen zur Informationskarte

1. Spaltenfeld (Sp. 1 - 4)
   Kartenkennung

2. Spaltenfeld (Sp 5 - 9)
   Jahrgangschiffre und Artikelnummer

3. Spaltenfeld (Sp. 10)
   Angabe, ob vollständiger Artikel oder Fragment

4. Spaltenfeld (Sp. 11 - 13)
   Publikationsform

5. Spaltenfeld (Sp. 14)
   Herkunft des Textes (Ost oder West)

6. Spaltenfeld (Sp. 15)
   Aufnahmeprinzip (1 = Modellmenge, 2ff = Zusätze)

7. Spaltenfeld (Sp. 16 - 26)
   Stellenangabe: Sigle des Textes (der Zeitung), Jahr, Monat, Tag, Seite

8. Spaltenfeld (Sp. 27 - 48)
   Angabe des Verfassers (sofern namentlich genannt) bzw. Ersatzformen solcher Angaben

9. Spaltenfeld (Sp. 49 - 57)
   Sigle(n) der Agentur(en), soweit erwähnt (bis zu drei nebeneinander)

10. Spaltenfeld (Sp. 58 - 63)
    Angabe des oder der Sachgebiete, in das (die) der Artikel gehört. Es können bis zu drei Sachgebiete angegeben werden. Die Reihenfolge soll dem Gewicht (Rang) entsprechen.
    Hauptsachgebiete: 1 = Politik, 2 = Wirtschaft, 3 = Soziales, 4 = Sport, 5 = Kulturelles

Gliederung der Sachgebiete

1	Politik
1A	Allgemeine Lage
1B	Persönlichkeiten
1C	Übernationale Zusammenschlüsse und Vereinbarungen
1D	West-Ost-Politik (bestimmter Staaten)
1E	Westliche Außenpolitik
1F	Östliche Außenpolitik
1G	Politik der Blockfreien und Neutralen (als Gruppen)
1H	Politik einzelner blockfreier und neutraler Staaten
1I	Militär- und Verteidigungspolitik (internat. u. national)
1J	Europäische Politik (auch einzelner Staaten)
1K	Afrikanische Politik " " "
1L	Asiatische Politik " " "
1M	Arabisch-nahöstl. Politik " " "
1N	Deutsche Außenpolitik
1O	Deutsche West-Ost-Politik (gesamtdeutsch)
1P	Innenpolitik auf Bundesebene bzw. der DDR
1Q	Innenpolitik auf Länderebene
1R	Innenpolitik auf Partei(en)ebene
1S	Deutsche Wirtschaftspolitik
1T	Innenpolitik best. Staaten
1U	Deutsche Kulturpolitik
1V	Deutsche Sozialpolitik
1W	Verbände
1X	Kommunalpolitik
1Y	
1Z	Sonstiges

2	Wirtschaft
2A	Allgemeine Lage
2B	Persönlichkeiten
2C	Technik
2D	Übernationale Zusammenschlüsse (außer europäischen)
2E	Europäische Zusammenschlüsse
2F	Organisationen, Verbände
2G	Planung
2H	Arbeitskräfte, Personalstruktur
2I	Industrie, Grundstoffindustrie
2J	Industrie, Investitionsgüter-
2K	Industrie, Konsumgüter-
2L	Landwirtschaft, Forsten
2M	Seefahrt, Fischerei
2N	Handwerk
2O	Handel, Außen-
2P	Handel, Innen- (auch gesamtdeutscher)
2Q	Betriebsorganisation, Rationalisierung
2R	Finanzen, Börse, Versicherungen
2S	Verkehr, Verkehrswege
2T	Verkehrsmittel (Luft-, Land-, Seefahrzeuge) und - technik
2U	Stadtplanung, Raumordnung
2V	Wohnungsbau, Bauwesen
2W	Dienstleistungsindustrie u. -gewerbe (Fremdenverkehr, Hotel-, Gaststätten u. dergl.)
2X	Werbung
2Y	Vergnügung, Unterhaltung, Massenmedien (nur wirtschaftlich)
2Z	Sonstiges
2(	Energiewirtschaft
2)	Handel-Versorgung

3	Soziales
3A	Allgemeines, Statistik
3B	Persönlichkeiten
3C	Bevölkerung
3D	Verbrechen, Vergehen (auch Beschuldigungen)
3E	Rechtswesen
3F	Gesundheitswesen, Medizin
3G	Verbände, Organisationen
3H	Körperpflege
3I	Versorgung (Renten usw.)
3J	Rassenprobleme
3K	Kommunikationsmittel (Presse, Rundfunk, Fernsehen, Film; aber nicht Artikel zum künstlerischen Rang!)
3L	Wissenschaft und Forschung
3M	Schul-, Erziehungs-, Bildungswesen
3N	Beruf und Arbeitswelt
3O	Haus und Garten
3P	Familie (Ehe)
3Q	Liebe (Ehe), Erotik (sog. Privat-oder Intimsphäre)
3R	Hobbies, Freizeit
3S	Geographisches
3T	Mode
3U	
3V	Veranstaltungen
3W	Naturereignisse, Wetter
3X	Unglücksfälle, Todesfälle
3Y	Streiks, Unruhen, Aufstände
3Z	Sonstiges

4	Sport
4A	Allgemeines
4B	Persönlichkeiten, Verbände
4C	
4D	Großveranstaltungen mehrerer Sparten (Olympiade, allgemein)
4E	Leichtathletik
4F	Schwerathletik, Boxen, Ringen
4G	
4H	Turnen, Gymnastik, Tanz
4I	
4J	Ballspiele, Mannschafts- (Fußball, Handball, Hockey usw.)
4K	
4L	Ballspiele, Einzel- (Tennis, Tischtennis usw.); Golf, Kricket
4M	Fechten
4N	Pferdesport
4O	Wassersport
4P	Wintersport mit Eislauf
4Q	Schießsport
4R	Radsport
4S	Motor- und Flugsport
4T	
4U	Bergsteigen, Wandern
4V	
4W	Spiele und Rätsel
4X	
4Y	
4Z	Sonstiges

5	Kulturelles
5A	Allgemeines
5B	Persönlichkeiten
5C	
5D	Organisationen, Gruppen
5E	kulturelle Entwicklung in einzelnen Ländern
5F	Musik (Konzert)
5G	Oper, Operette, Musical
5H	Theater, Ballett
5I	Malerei, Graphik
5J	Bildhauerei, Plastik
5K	Architektur
5L	Literatur
5M	Sprache
5N	Wissenschaft, Forschung, Universitäten (nur zu deren kulturellen Aufgaben und Leistungen, nicht zur sozialen Funktion)
5O	Massenmedien (Film, Rundfunk, Fernsehen)
5P	Philosophie, Weltanschauung, Ideologie
5Q	
5R	Religion und Kirche
5S	
5T	Geschichte
5U	
5V	Volks- und Brauchtum
5W	
5X	
5Y	
5Z	Sonstiges

11. Spaltenfeld (Sp. 64-70)

Innere und äußere Form

a) Innere Form

1) Publizistische Ziele (Sp. 64-65)

Es können 2 Chiffren gleichzeitig angegeben werden.

1 = Unterrichtung
2 = Belehrung
3 = Beeinflussung
4 = Unterhaltung
5 = Wirtschaftswerbung

2) Mitteilungsform (Sp. 66)

A = Nachricht (Unterrichtung, Wirtschaftswerbung)
B = Bericht (Unterrichtung, Belehrung, Beeinflussung, Wirtschaftswerbung)
C = Background-Bericht (Unterrichtung, Belehrung, Beeinflussung)
D = Abhandlung (Belehrung, Unterrichtung)
E = Beitrag (Belehrung, Unterrichtung, Unterhaltung, Wirtschaftswerbung)
F = Tips, Voschlag, Anweisung (Belehrung, Unterrichtung Unterhaltung, Wirtschaftswerbung)
G = Hauptkommentar (Beeinflussung, Belehrung, Unterrichtung)
H = Glosse (Beeinflussung, Unterhaltung)
I = "Spitzmarke" (Beeinflussung, Unterhaltung)
J = Geschichte (Unterhaltung, Belehrung) (Kurzgeschichte, Erzählung)

K = Plauderei (Unterhaltung, Beeinflussung, Belehrung)

L = "kleines Feuilleton" (Unterhaltung, Beeinflussung, Belehrung)

M = Großanzeige (Wirtschaftswerbung, Beeinflussung, Unterrichtung, Belehrung)

N = Anzeige (Wirtschaftswerbung, Beeinflussung, Unterrichtung, Belehrung)

Sonderformen

O = Text zu einem Bild, Schaubild, graph. Darstellung usw.

P = Leserbrief

Q = Interview

R = Abdruck eines literarischen Dokuments (auch Gedichtes)

S = Abdruck einer amtlichen Verlautbarung (Gesetz, Vertrag, Bekanntmachung, Aufruf), Kommuniqué

T = Abdruck einer Rede

U = Abdruck eines fremden Artikels

V = Klein- und Privatannonce

W = Tabelle, Wetterbericht

X = Fortsetzungsroman

Z = Sonstiges

Für diese Sonderformen braucht kein publizistisches Ziel angegeben zu werden; die dafür vorgesehenen Spalten (64-65) werden dann mit 00 ausgefüllt.

3) Angabe, ob es sich um eine Übersetzung handelt (Sp. 67)

4) Angabe der Sparte (Sp. 68)

Politik	= 1
Wirtschaft	= 2
Sport	= 3
Technik-Motor	= 4

Feuilleton = 5
Lokales = 6
Anzeigen = 7
Vermischtes = 8
Sonstiges = 0

In Zeitungen, die selbst keine Sparten angeben (meist am oberen Rand der Seite), kann diese Spalte zunächst frei bleiben.

b) Äußere Form (Sp. 69-70)
   1) Länge des Artikels in Zeilen (nach Klassen gruppiert),
   2) Breite des Artikels in Spalten.

12. Spaltenfeld (Sp. 71-77)
    Angaben über stilistische Kriterien.
    (noch nicht entwickelt!)

13. Spaltenfeld (Sp. 78-80)
    Länge des Artikels in Karten.

Erläuterungen zur Beikarte (Themakarte)

1. Spaltenfeld (Sp. 1-4)
   Kartenkennung
2. Spaltenfeld (Sp. 5-9)
   Jahrgangschiffre und Artikelnummer
3. Spaltenfeld (Sp. 10)
   bleibt leer
4. Spaltenfeld (Sp. 11-80) Themaangabe
   a) Kopfteil:
      enthält Angabe des oder der im Artikel behandelten Landes (Länder).
      Bundesrepublik = BRD, DDR = DDR, für alle anderen Staaten die Chiffren der amtlichen Autokennzeichen.
      Es können bis zu vier Staaten im Kopfteil genannt werden.

Danach ist zweimal blank zu setzen. Bezieht sich der Artikel auf kein bestimmtes Land, ist nach Spalte 10 zweimal blank zu setzen.

b) Textteil: Stichwort-Angabe des behandelten Themas.

ANHANG IV

Übersicht über die in der Außenstelle vorhandenen wichtigeren Programme

Nr.	Name Gebiet	Leistung
1	PRUEF	("Textprüfung")
	Texterstellung	Übernimmt Daten von Karten oder Band auf Band, prüft Textfiles auf numerische Richtigkeit nach folgenden Kriterien: 1. lückenlose, wiederholungslose, aufsteigende Folge von a) Artikeln, b) Karten (vergleicht beide Numerierungen), 2. Vorhandensein, Lage und Zugehörigkeit von Informationskarten, 3. Beginn und Ende der Artikel nach besonderen Bedingungen. Gibt 20 verschiedene Kommentare als Prüfprotokoll aus. Zusätzlich: Erstellt Textausdrucke in verschiedener Form mit Kommentierung.
	Unterprogr.	EINAUS (IPK), 4 weitere für Ausdruck, 5 für Prüfung.
	Autor	W. Klutentreter, IDS
2	VORKOR	("Vorkorrektur")
	Texterstellung	Übernimmt zur Bandkorrektur bestimmte Korrekturkarten auf Band, prüft auf Lage, Format und Erfüllung der von KORR (Nr.3) verlangten Eigenschaften. Gibt 13 Kommentare als Prüfprotokoll aus.
	Unterprogr.	3 zum Prüfen
	Autor	W. Klutentreter, IDS
3	KORR	("Korrektur")
	Korrektur beliebiger files hier: Texterstellung	korrigiert files (hier: Textfiles) anhand von Korrekturkarten durchErsatz, Löschen oder Neueinfügen einzelner Records (bei Löschen u. Neueinfügen s.Nr.4)
	Unterprogr.	EINAUS, CHAR, BIST, ZELE
	Autor	T. Krumnack, IPK

Nr.	Name Gebiet	Leistung
4	NDNUM	("ND-Numerierung")
	Texterstellung	Numeriert lückenhafte oder überbelegte files (oder Teile von files) neu durch - hier: von KORR korrigierte Textfiles.
	Unterprogr.	-
	Autor	H. E. Neuhaus, IPK
5	INKASU	("Informationskarten-Suche")
	Textbereit-stellung	Sucht nach vorgegebenen Kriterien (auch Kombinationen mehrerer Kriterien) Informations- und Beikarten ab, schreibt Nummern der betr. Artikel aus und/ oder stellt die betr. Artikeltexte auf Band zur Verfügung oder druckt aus.
	Unterprogr.	
	Autor	G. Billmeier, IDS
6	MOSU 1	("Morphem-Suche" 1. Teil)
	Wort-Suche	Sucht auf Textbändern aufgrund vorgegebener Buchstabenkombinationen Belegwörter und übergibt sie mit Kontext und Stellenangaben sowie Informationskarten an Folgeprogramm. (Kann auch als Unterprogramm verwendet werden) Spezifikationen: a) Suche ist beschränkbar auf "Endungen", b) jede Stelle der Such-Kombination kann durch mehrere Zeichen belegt werden, c) Substantive oder Nicht-Substantive können ausgefiltert werden, d) reiner Zähllauf (ohne Belegausgabe) ist möglich, e) Kontext kann von 1-11 Karten variiert werden.
	Unterprogr.	EINAUS, SIEB, ZELE, ZAHL
	Autor	T. Krumnack, IPK
7	MOSU 2	("Morphem-Suche" 2. Teil)
	Wort-Suche	Übernimmt von MOSU 1 gefundene Belegwörter mit Kontext, sortiert sie nach Such-Kombinationen (immer) und - je nach Spezifikation - innerhalb dieser alphabetisch, reinigt Kontext von nicht gewünschten Zeichen, gibt mit Hilfe von INFDEC (Nr.8) Informationen

Nr.	Name Gebiet	Leistung
		aus Informationskarte bei, beschneidet Kontext auf gewünschte Länge (Zahl von Karten, Zahl von Wörtern oder satzweise) und druckt aus. Liegt in 3 Standardvarianten, 2 nicht standardisierten Varianten und 1 erweiterten Variante vor:
	a) SAMOS	("Standard-Ausdruck für MOSU") Liefert: gefundenes Wort, Fundstellenangabe von Inform.-karte (falls gewünscht) und Kontext (mit Nummer) in folgenden drei Stufen: SAMOS 1: 3 Zeilen Kontext (= Belegzeile, Vorzeile, Folgezeile), sortiert alphabetisch bis 400 Belege je Suchkombination; SAMOS 2: 5 Zeilen Kontext; sortiert bis 250 Belege je Suchkombination; SAMOS 3: 7 Zeilen Kontext; (sortiert Belege nicht innerhalb der Suchkombination).
	b) SAMOS W	Wie SAMOS, aber Kontextbeschränkung nach angebbarer Zahl von Wörtern; Belege unbeschränkt sortierbar;
	c) SAMOS S	Wie oben, aber Kontextbeschränkung auf 1 Satz (von Satzschlußzeichen zu Satzschlußzeichen) (Belege nicht sortiert; begrenzte Kapazität);
	d) SAMOS E	Kombination von SAMOS W und INFDEC. Wie SAMOS W, decodiert zusätzlich bestimmte gewünschte Angaben aus Informationskarten und fügt sie den Belegen bei; erstellt nicht automatisch Druckspiegel.
	Unterprogr.	
	Autor	G. Billmeier, IDS
8	INFDEC	("Informationskarten-Dechiffrierung")
	Spezial-Unterprogramm	Dechiffriert angegebene Bereiche der Informationskarten in Klartext und übergibt an jeweiliges Hauptprogramm, vor allem: an PRUEF (Klartext der Inform.-karten am Beginn jede Artikels in Klartextausdrucken), an MOSU (SAMOS E) zur näheren Kennzeichnung der Belege. Mit eigenem Hauptprogramm: Erstellt dechiffrierte Wiedergabe der Inform.-karten in Karteikartenformat zur Kontrolle der Inform.-karten.

Nr.	Name Gebiet	Leistung
	Unterprogr.	-
	Autor	G. Billmeier, IDS
9	SUSA	("Suche von Stern-Abkürzungen")
	Wort-Suche	Sucht auf Textband Initial-Abkürzungen (= Wörter mit mehr als 1 "Stern" (Großschreibungszeichen)) mit Belegstellen. Auch als Unterprogr. zu SUBEL verwendbar.
	Unterprogr.	EINAUS, PACK
	Autor	H. D. Lutz, IDS
10	SUBEL	("Suche von Belegstellen")
	Wort-Suche	Sucht anhand vorgegebener (auch von SUSA übernommener) Belegstellenangaben auf Textband Kontext variablen Umfangs und druckt aus.
	Unterprogr.	EINAUS, PACK, (SUSA)
	Autor	H. D. Lutz, IDS
11	SUBIN	("Suche von Bindestrich-Komposita")
	Wort-Suche (Spezial-Progr.)	Sucht auf Textband Bindestrich-Komposita, klassifiziert sie nach 6 Gesichtspunkten, druckt Belege mit variablem Kontext aus.
	Unterprogr.	EINAUS, 5 interne
	Autor	G. Billmeier, IDS
12	COSYMA	("Convertierungs-System f. Mannheimer Bänder")
	Textbereit-stellung	Wandelt Band-Texte nach Mannheimer Schreibkonventionen in Band-Texte nach Bonner Konventionen um.
	Unterprogr.	EINAUS
	Autor	G. Billmeier, IDS

Nr.	Name Gebiet	Leistung
13	COSYBO	("Convertierungs-System f. Bonner Bänder")
	Textbereit-stellung	Wandelt Band-Texte nach Bonner Schreibkonventionen in Band-Texte nach Mannheimer Konventionen um.
	Unterprogr.	EINAUS
	Autor	G. Billmeier, IDS
14	DECOS	("Decodierung von Streifentexten")
	Textbereit-stellung (Lochstreifen)	Übernimmt Text von Lochstreifen (hier: TTS-Zeitungs-setzstreifen), codiert Zeichen um, säubert von unerwünschten Zeichen, gleicht Text z.T. an Bonner Schreibkonventionen an, bringt Text in Records bestimmter Länge, definiert und numeriert Artikel und Zeilen, erkennt Überschriften, druckt aus oder übergibt an ESLOT.
	Unterprogr.	
	Autor	W. Klutentreter, IDS
15	ESLOT	("Erkennung v. Satzschlüssen in Lochstreifen-Texten")
	Textbereit-stellung (Lochstreifen)	Übernimmt durch DECOS umcodierte und vorbereitete Streifentexte, unterscheidet Satzschlußpunkte von anderen Punkten, definiert Satzschlüsse, schreibt Satzanfänge klein (sofern nicht Substantiv am Satzanfang), kommentiert Entscheidungen und druckt aus.
	Unterprogr.	EINAUS, 6 - 7 interne
	Autor	G. Billmeier, IDS
16	IDVS	("IDS-Vorsortier-Programm")
	Index-Erstellung	Normiert Texte im IDS-Code (Bonn) zu sortierbaren Einheiten, getrennt nach Einzelwörtern und Wort-gruppen (Transkriptionsgruppen und Bindestrich-Gruppen).
	Unterprogr.	EINAUS, IDWO, IDTA u. a.
	Autor	T. Krumnack, IPK
17	IBSORT	("IBSYS-Sortier-Programm")
	Index-Erstellung	Sortiert nach IDVS genormte Band-files alphabetisch, getrennt nach Einzelwörtern und Wortgruppen. Oder: Sortiert von RELI (Nr. 19) erstellte files nach relativen Häufigkeiten.

Nr.	Name Gebiet	Leistung
	Unterprogr.	s. IBSYS-Systemband
	Autor	T. Krumnack, IPK
18	IDMI	("IDS-Misch-Programm")
	Index- Erstellung	Mischt von IBSORT getrennt sortierte Einzelwörter und Gruppen in alphabetischer Reihenfolge.
	Unterprogr.	EINAUS, BIST, ZELE, ZAHL
	Autor	T. Krumnack, IPK
19	RELI	("Listen mit relativen Häufigkeiten")
	Index- Erstellung	Erzeugt aus von IDMI oder IBSORT sortierten files alphabetische Indices mit Angaben der Fundstellen, der absoluten und relativen Häufigkeiten sowie der Zugehörigkeit zu einer Wortgruppe. Druckt aus oder übergibt an IBSORT.
	Unterprogr.	EINAUS, SIEB, BIST, ZELE, ZAHL
	Autor	T. Krumnack, IPK
20	KRLI	("Listen nachKlassen relativer Häufigkeit")
	Index- Erstellung	Übernimmt files, die nach Häufigkeit sortiert sind (und innerhalb gleicher Häufigkeit alphabetisch), a) listet aus: Häufigkeitsklasse, Zähler pro Klasse, relative Häufigkeit, absolute Häufigkeit, Wort; b) gibt a) auf Band, zusätzlich mit Kennung, die den file bezeichnet, zwecks Weiterverarbeitung durch VERSYS (Nr. 21).
	Unterprogr.	
	Autor	T. Krumnack, IPK
21	VERSYS	("Vergleichssystem für Häufigkeitsfiles")
	Vergleichs- register	a) Vergleicht zwei alphabetisch geordnete files (2 Bänder), erstellt daraus allgemeines Vergleichsband (mit Angabe zu jeder Wortform, ob vorhand und wie häufig (nach Klassen)). b) Erzeugt aus Vergleichsband spezielle Vergleichs-Listen (intern vorgesehen: 20 mögliche Listen). c) Vergleicht zwei alphabetisch geordnete files auf einem Band, erstellt daraus Listen nach b).

Nr.	Name Gebiete	Leistung
		Es können Vergleichs-Register erzeugt werden nach: a) "linker" Differenz b) "rechter" Differenz c) "Durchschnitt" (Gemeinsames) d) Vereinigung (aber nach links und rechts getrennt). Vergleich kann erfolgen nach 1. vorhanden - nicht vorhanden 2. gleich - ungleich 3. häufig - selten (nach wählbarer Definition bzw. Gewichtung).
	Unterprogr.	
	Autor	H. E. Neuhaus, IPK

Anmerkungen

zum Forschungsbericht Hellmann (Außenstelle Bonn)

1) In begrenztem Umfang sind Untersuchungen zur gesprochenen Sprache möglich. In der Außenstelle wurde einige Monate lang von der damaligen wiss. Mitarbeiterin Frau Dr. Inge Kraft eine Reihe von Tonbandaufnahmen aus Landfunksendungen des Senders "Radio DDR", und zwar Interviews und Gespräche, untersucht und mit westlichen Landfunkinterviews sowie mit östlichen und westlichen Zeitungsartikeln über landwirtschaftliche Themen verglichen. Die in den Interviews von Radio DDR gesprochene Sprache zeigte, von einigen charakteristischen Fehlleistungen abgesehen, eine so weitgehende Übereinstimmung mit der Terminologie gedruckter Texte (Zeitungen), d.h. eine so starke Anpassung des Interviewten an die Redeweise des jeweiligen Interviewers, daß die Untersuchungen wieder eingestellt wurden. Frau Dr. Kraft hat während der Herbsttagung 1965 des Instituts in Mannheim über ihre Ergebnisse berichtet; ein Aufsatz wird in Kürze in "Deutsche Studien" Heft 23, Lüneburg 1968, erscheinen.

2) Zum Problem der Vergleichbarkeit unseres westlichen und östlichen Materials vgl.unten Punkt 7.3.

3) Dazu vgl. Anhang III, Abs. 11a (4).

4) Zur Berücksichtigung der in unserem Material vorgefundenen Sachgebiete, s. Anhang III, Abs. 10.

5) Der Begriff "Thema" ist nicht mit "Sachgebiet" zu verwechseln. Zur Kennzeichnung vgl. unten Punkt 13.2.3.

6) Gewisse Themen kehren allerdings in bestimmten Abständen wieder; dazu unten 4.5.2.1.b).

7) Zur Definition des Begriffs s.4.6.1. und 13.1.

8) E. Dovifat, Zeitungslehre Bd. 1, Berlin 1962 (Göschen H. 1039) unterscheidet nur "Nachrichtenstilform, Meinungsstilform, Unterhaltungsstilform"(S. 120, Näheres S. 120 - 133).
W. Hagemann, Die Zeitung als Organismus, Heidelberg 1950, unterscheidet "Unterrichtung, Beeinflussung, Wirtschaftswerbung, Unterhaltung, Belehrung" (vgl. a.a.O. Kap. III, S.35 ff). Beide ordnen den publizistischen Zielen bestimmte Schreib- oder Stilformen zu. - Zur Anwendung dieser Kriterien auf unser Material vgl. Anhang III, Abs. 11a (1 und 2).

9) Vgl. dazu Anhang III, Abs. 11a (Innere Form). Unsere Einteilung lehnt sich an die Hagemanns an (a.a.O. S. 49 ff.).

10) Zwingend ist dieser Schluß freilich nicht; theoretisch ist es möglich, auch eine Vielzahl von Sachgebieten, Themen usw. mit relativ sparsamen sprachlichen Mitteln zu bewältigen.

11) Daraus den Schluß zu ziehen, man habe es bei Tageszeitungen sicher mit "Normal-" oder "Durchschnittssprache" zu tun, ist allerdings gefährlich; hier können zwei Fehlerquellen liegen:
   a) Die Redaktion selbst kann sich in ihrer Leserschaft irren, sie kann sie über- und unterschätzen, wenn sie sich z.B. einer Redeweise bedient, die bei der Leserschaft oder einem Teil davon "nicht ankommt".
   b) Der (wiss.) Betrachter kann sich irren, nämlich wenn eine Redaktion absichtlich, z.B. um eine propagandistische Wirkung nach außen zu erzielen, sich einer Redeweise bedient, die den Eindruck einer besonders fortschrittlichen (oder kämpferischen, sozialistischen, national gesinnten, christlichen, gebildeten, aufgeschlossenen usw.) Leserschaft erwecken soll. Dies dürfte aber wohl nur in einer Gesellschaft möglich sein, in der ein Meinungsmonopol herrscht. Im übrigen bleibt auch dann die Frage offen, inwiefern nicht auch eine solche in bestimmter Richtung forcierte Redeweise schließlich Wirkungen auf die Redeweise der Leser hat, gerade unter den Bedingungen eines Meinungsmonopols.

12) Unter "Dokumentation" wird hier nicht erst die (maschinelle) Aufbereitung von Texten, ihre Ausstattung mit ständig zu erweiternden Informationen zwecks (maschineller) Reproduktion sprachlicher Daten unter verschiedensten Fragestellungen verstanden, sondern auch schon die diesem Ziel dienende adäquate Ermittlung der entsprechenden Texte.

13) Ohne Zweifel hat diese und jede andere formale Einteilung keine unmittelbare qualitative Bedeutung. Eine Ausgabe vom 31. Dezember ist von der des 1. oder 2. Januar nicht grundsätzlich stärker unterschieden als von der des 30. Dezember.

14) Sie ist insofern auch berechtigt, als sich in jedem Jahrgang bestimmte Themen turnusmäßig, nämlich jahreszeitlich bedingt, wiederholen. Dazu vgl. unten 4.5.2.1.b).

15) Zum Vergleich die Zahlen für das "Neue Deutschland" des Jahrgangs 1964: Gesamtmenge ohne Sonntagsausgaben und Beilagen: ca. 2200 Seiten entspr. 7 Mill. lfd. Wörter (dabei ist berücksichtigt, daß 1 Seite ND durchschnittlich mehr Wörter enthält als 1 Seite WELT, da dort großräumige Anzeigen sehr selten sind). Diese Menge entspricht rund 14.000 Buchseiten.

16) ND Februar 1964: 184 Seiten ohne Beilagen und ohne Sonntagsausgaben; eine Zeitungsseite ND entspricht etwa 6-7 Buchseiten.

17) Die erste und trivialste dieser Schwierigkeiten besteht darin, daß die fast ausschließlich benutzten TTS-Streifen mit den in den Rechenzentren und wiss. Instituten vorhandenen Streifenlesegeräten nicht zu verarbeiten sind. Lesegeräte, die sowohl Standard- als auch TTS-Streifen lesen können, gibt es zwar, sie werden aber an den für uns erreichbaren Stellen nicht eingesetzt. Am Germanistischen Seminar der Universität Saarbrücken besteht zwar die Möglichkeit, TTS-Streifen in Standardstreifen umzuformen, jedoch ist dies ein umständlicher Weg, zumal bei regelmäßiger Umformung von Streifen in großen Mengen.
Eine weitere Schwierigkeit liegt darin, 1. daß die TTS-Streifen durchweg nicht korrigiert sind, 2. daß Überschriften, Anzeigen sowie andere Texte mit drucktechnischen Besonderheiten nicht über Streifen gesetzt werden und umgekehrt nicht alles, was auf den Streifen steht, auch in der Zeitung erscheint (größere Änderungen noch kurz vor dem Druck).
Ferner ist die Mehrdeutigkeit einiger Zeichen meist unerwünscht, wie z.B. Nichtunterscheidung von Abkürzungspunkt und Satzschlußpunkt, Großschreibung am Satzanfang und damit Nichtunterscheidung von der sonst willkommenen Großschreibung der Substantive usw. In der Außenstelle ist inzwischen ein Programm erarbeitet worden, das diese Schwierigkeiten weitgehend automatisch beseitigt. Es arbeitet z. Zt. mit einer Fehlerquote von knapp 5 Prozent. (Weiteres vgl. Anhang IV "Programmübersicht", Nr. 14 u. 15). Schließlich werden von elektronisch gespeichertem Material durchweg zusätzliche Informationen sprachlicher Art erwartet, die ein über Streifen eingelesener Text ohne Eingriffe von Hand natürlich nicht enthält. Dennoch erscheint es nach den bisher erreichten Ergebnissen als möglich, durch geeignete Aufbereitungsprogramme auch fremde Streifentexte so zu verbessern, daß sie sich als Textgrundlage für unsere Zwecke eignen, sofern man die Forderung nach Eindeutigkeit und Zusatzinformationen etwas reduziert.

18) Vgl. G. Herdan, The Advanced Theory of Language as Choice an Chance, Berlin-Heidelberg-New York 1966, S. 96 f. und 169 ff. .

19) Die Begriffe "repräsentative Auswahl bzw. Menge" und "Repräsentanz" bzw. "Repräsentativität" werden bewußt vermieden, da sie, wenigstens im allgemeinen Sprachgebrauch, die Vorstellung des Absoluten, fest Definierbaren auslösen können. Es gibt jedoch in diesem Sinne keine absolut repräsentative Auswahl.

20) Der Begriff "Modell" gibt das Gemeinte insofern adäquat wieder, als er die Vorstellung intendiert, daß einModell maßstabgetreu und funktionsgerecht zu sein hat, dies jedoch in Einzelheiten nur dann sein kann, wenn ein bestimmter Maßstab, der von den jeweils an das Modell zu stellenden Anforderungen abhängt, nicht unterschritten wird. Dabei ist klar, daß auch das beste Modell nur einen Teil der Eigenschaften der Bezugsgröße genau, viele andere nur annäherungsweise und einige gar nicht widerspiegelt.

21) Zur Unterscheidung von Gesamtmenge und Zielmenge vgl. auch 4.5.3.1.a).

22) Vgl. 4.5.2.1. und 4.5.3.1.

23) E. Mittelberg, Wortschatz und Syntax der Bildzeitung, Marburg 1967, hat für seinen Untersuchungsgegenstand gezeigt, daß dort die herkömmliche Sparteneinteilung und Aufeinanderfolge der Sparten nicht gilt. Für die Organe bestimmter Institutionen oder Gruppen (Verbände usw.) ist ähnliches anzunehmen.

24) Die Dokumentation von Zeitungstexten steht hier vor ähnlichen Problemen wie die "repräsentativen Bevölkerungsumfragen": auch dort können alle nach fein gegliederten Kriterien wie Geschlecht, Alter, Familienstand, Schulbildung, Konfession, Beruf, Wohnort usw. vorgenommenen Befragungen zu falschen Ergebnissen führen, wenn tatsächlich nicht eines der berücksichtigten Kriterien, sondern ein nicht berücksichtigtes wie etwa die Programmwahl im Fernsehen den Ausschlag für das erwartete Verhalten der Bevölkerung gibt.

25) So kann etwa die scheinbar mögliche Unterscheidung von namentlich gekennzeichneten und anonymen Artikeln kein echtes Kriterium einer Auswahl sein, weil die Zeitungen hier sehr willkürlich verfahren.

26) Die Verwendung der Einheit "Seite" als Maß- oder Zähleinheit ist nicht zu verwechseln mit ihrer möglichen Verwendung als Aufnahmeeinheit. Dazu s. unten 4.6.1.

27) Sie gilt für Druckerzeugnisse jeglicher Art, also außer für Bücher und Periodika auch für Sonderdrucke, Flugschriften, Manuskripte, Plakate usw.

28) Beispiele für die Vielfalt der "Beilagen" allein in der WELT: Die Rubriken "Betrieb und Beruf" sowie "Schule und Hochschule" erscheinen nicht täglich, sind zwar im allgemeinen laufend durchnumeriert, nehmen aber doch nach Umfang und Aufmachung den Charakter von Beilagen an. Regelmäßig erscheinende Beilagen sind etwa die Stellenanzeigen in den Samstagausgaben (obwohl auch im übrigen durchnumerierten Teil der Zeitung noch Stellenanzeigen stehen können); ferner "Die geistige Welt" (ebenfalls samstags) und die "Reise-Welt" (freitags); diese Beilagen sind durchweg, jedoch nicht immer, gesondert (römisch) numeriert. Das schließt aber nicht aus, daß nicht auch in anderen Ausgaben mehrere Seiten für "Reise und Erholung" reserviert werden, ohne direkt Beilagen zu sein. - Die "Welt der Literatur" nimmt eine Sonderstellung schon im Format ein (halbes WELT-Format); sie erscheint 14-tägig donnerstags und ist gesondert abonnierbar, stellt also fast eine Zeitung in der Zeitung dar.
Anläßlich der Olympischen Spiele haben mehrere Zeitungen Sonderbeilagen herausgebracht.

29) Nach dieser Definition kann die "Ausgabe" also auch nur etwa die Beilagen einer Zeitung enthalten. Die Summe aller Erscheinungsmengen ist also die Gesamtmenge, die Summe aller Ausgaben ist die Zielmenge.

30) Falls also eine Zeitung im Laufe eines Jahrgangs von wöchentlichem zu täglichem Erscheinen übergeht (oder umgekehrt), können die folgenden Überlegungen nicht angewandt werden.

31) Für ein 1-bändiges Buch ergibt sich der Wert 365, da n (Zahl der Ausgaben im Jahr) gleich 1 ist.

32) Im Jahrgang 1966 des "Neuen Deutschland", das täglich erschien, sind 359 Nummern herausgekommen, also 6 Nummern ausgefallen; die WELT hat 1966 305 Nummern herausgebracht, d.h. nach Abzug der 52 Sonntage sind 8 Nummern ausgefallen. Die ausgefallenen Nummern betreffen sämtlich die gesetzlichen Feiertage.

33) Erfahrungsgemäß reicht es aus, etwa jede vierte Woche oder jeden vierten Monat eines Jahrgangs, also ca. 1/4 der Jahrgangsmenge, genau auszuzählen. Größere Genauigkeit ist nicht erforderlich, da sie in der Auswahl ohnehin nicht berücksichtigt werden könnte.

34) Bei der Ermittlung unserer Auswahl wurden jeweils 4 bzw. 5 Monate genau ausgezählt (vgl. unten Punkt 10.1., Anm. 50 und 51, und Anhang Ia).

35) Wieviel Aufnahmeeinheiten (Seiten, Artikel) eine Stichprobeneinheit enthält, hängt von der Menge ab, die der Bearbeiter aufzunehmen für notwendig hält, d.h. von der gewünschten oder geforderten Aufnahmequote bzw. dem Wahrscheinlichkeitsquotienten.

36) Die untere Grenze für das Aufnahmeintervall ist selbstverständlich das Erscheinungsintervall.

37) Konkret bedeutet das: In einer auf einen ganzen Jahrgang bezogenen Modellmenge sind 10 ganze Ausgaben à 10 Seiten selbstverständlich besser als 10 einzelne Seiten, aber 100 gut verteilte einzelne Seiten sind wesentlich besser als 10 ganze Ausgaben.

38) Näheres s. Teil III, S 140 ff.

39) Läßt man also eine Auswahl am 10. März beginnen und führt sie mit einem Aufnahmeintervall von 6 Tagen bis zum 10. Juni fort, so kann diese Auswahl selbst bei genauester Beachtung aller Bedingungen einer zureichenden Auswahl genau für den Zeitraum vom 7. März bis 13. Juni Modell sein, jedoch niemals für den ganzen Jahrgang oder für einen anderen Zeitraum.

Damit ist nicht gesagt, daß nicht auch eine zeitlich sehr eng begrenzte Auswahl sinnvoll sein kann. Für eine Untersuchung etwa, für die ein diachronisch extrem tiefgestaffeltes Material benötigt wird (etwa bei eng begrenzten Wortschatzuntersuchungen über 10 bis 20 Jahrgänge hinweg) kann eine Auswahl aus nur einer Woche (allerdings immer der gleichen Woche) bei genügender Dichte eine sehr ergiebige Modellmenge darstellen; - freilich ist sie Modell für die Verhältnisse in dieser einen bestimmten Woche in diesen 10 bis 20 Jahrgängen. Bedingung ist dabei, daß man sich der zeitraumbedingten Besonderheiten gerade dieser kurzen Auswahlspanne genau bewußt ist. (Als Beispiel für sinnvolle Verwendung einer solchen zeitlich begrenzten Auswahl vgl. H. Bartholmes, Das Wort "Volk" im Sprachgebrauch der SED, Düsseldorf 1964).- Es sind also auch diachronisch gegliederte Modellmengen möglich, für die aber die aufgezeigten Grundsätze für die Ermittlung einer zureichenden Auswahl - entsprechend modifiziert angewandt - ebenfalls gelten.

40) Die noch verbleibende Korrekturarbeit - zu Beginn der Textaufnahme der überwiegende Teil - wurde von einem wissenschaftlichen Mitarbeiter übernommen.

41) Zu den Schreibkonventionen vgl. unten Punkt 13.

42) Diese Kapazität ist jedoch nur mit eingearbeiteten Kräften bei fertig entwickeltem System zu erreichen. Die tatsächlich erreichte Leistung lag im ersten Jahr bei etwa einem Viertel, im zweiten Jahr bei etwa der Hälfte der angegebenen Kapazität. Im Jahre 1967 ist die Kapazität nahezu erreicht worden.

43) Es wurde die Berliner Ausgabe (B-Ausgabe) statt der Republik-Ausgabe (A) gewählt, einmal um auch Texte mit (mehr oder weniger) regionalem Kolorit zu erhalten, vor allem aber weil auch die WELT in einer Berliner Regionalausgabe erscheint; die beiden Zeitungen wurden dadurch besser vergleichbar.

44) Das ND 54 hatte noch keine Trennung zwischen Berliner und Republik-Ausgabe.

45) So ist z.B. im ND ein Feuilleton im Sinne von "Unterhaltung" so gut wie unbekannt; ein kulturell-literarischer Teil erscheint sporadisch und mit sehr wechselndem Umfang, sofern er nicht als Beilage erscheint (dazu siehe unten 8.1.). Einen Roman "unter dem Strich" enthielt der Jahrgang 1964 nicht, dagegen ein Teil des Jahrgangs 1954.

46) Auflagenhöhe der drei Zeitungen (nach Ausgaben des Bundespresse- und Informationsamtes) für 1967:
FAZ: 249.00, WELT: 247.000, ND: 600.000.
Die Angaben für die Auflagehöhe des ND differieren, es handelt sich um die wahrscheinlich annähernd zutreffende Zahl. Dabei muß noch berücksichtigt werden, daß bei einem überwiegend propagandistisch orientierten Organ wie dem ND der Anteil der nur verteilten, nicht gelesenen Exemplare verhältnismäßig hoch liegt.

47) Vgl. zu den Mannheimer Zeitungstexten Teil I, Textliste (1.1.e).

48) Die Herstellung solcher Mischtexte kann unter bestimmten Bedingungen sinnvoll sein, etwa dann, wenn es darum geht, individualstilistische Eigenheiten in mehreren Texten sonst großer Ähnlichkeit zu unterdrücken (in der linguistischen Forschungsstelle in Besançon ist man z.T. diesen Weg gegangen). In diesem Sinne stellen aber unsere Zeitungen selbst schon hochgradige Mischtexte dar. Eine Mischung höherer Stufe würde die Untersuchung jedenfalls nicht erleichtern.

49) Unvollständige Wörter am Ende eines Artikelfragments werden ergänzt. Einige Artikelfragmente aus Ulbricht-Reden wurden ergänzt; vgl. Punkt 12. Verweismöglichkeiten bei nachträglicher Aufnahme fehlender Artikelteile s. 13.2.3.("Zusatzkarte").

50) Der Umfang der Zielmenge für ND 64 wurde ermittelt aufgrund der Auszählung von 5 Monaten (vgl. Anhang Ia/1), für ND 54 durch Auszählung von 4 Monaten (vgl. Anhang Ia/3).

51 Der Umfang der Zielmenge wurde ermittelt aufgrund der Auszählung von 5 Monaten (vgl. Anhang Ia,2).

52) Tatsächlich schwankt der Umfang der Stichproben zwischen 2 und 4 Seiten (s. Liste der aufgenommenen Seiten, Anhang IIa).

53) In der Praxis bedeutet das, daß Seiten aus verschiedenen Ausgaben gegeneinander ausgetauscht wurden, was besonders bei Seiten mit hohen Seitennummern, d.h. bei Seiten, die selten vorkommen, mehrfach nötig wurde.

54) Die aus dem ND 64 und 54 zur Aufnahme vorgesehenen Seiten erhielten wir freundlicherweise in Form von Mikrofilm-Rückvergrößerungen vom Archiv für gesamtdeutsche Fragen, Bonn; die aus WELT 64 vorgesehenen Seiten vom Archiv des Seminars für Publizistik an der Freien Universität Berlin, und zwar als Mikrofilm; auch hier nahm das Archiv für gesamtdeutsche Fragen die Rückvergrößerungen vor (s. auch Vorbemerkung S.8).

55) Diese Februar-Menge lag den Untersuchungen von G. Billmeier im Teil III zugrunde.

56) Die Artikel liegen in unserem Material und auf dem Magnetband in fortlaufender, aufsteigender Reihenfolge. Gelegentlich sind Nummern unbesetzt, da einige Male zunächst getrennt numerierte Artikel unter einer Nummer zusammengefaßt wurden; in anderen Fällen wurden Artikel nachträglich getrennt, dann erhielt einer der beiden Artikel eine der bisher freigebliebenen Nummern. Die Numerierung entspricht also nicht vollständig der Artikelfolge auf den Zeitungsseiten. Die Zuordnung von materialinterner Artikelnummer zum Original-Zeitungsartikel ist mit Hilfe der Informationskarte dennoch gesichert (vgl. Anhang III, 1.Seite).

57) Vgl. dazu die Mannheimer Schreibanweisungen, Teil I, Punkt 2.4.

58) Näheres zu den einzelnen Stufen des Verfahrens unten in den "Erläuterungen zum Arbeitsablauf", Punkt 14.1.

59) Vgl. dazu unten 13.2.2. und Anm. 67.

60) Diese Ersatzzeichen sowie auch andere Bestandteile unserer Schreibkonventionen gehen auf die Schreibkonventionen des Instituts für Phonetik und Kommunukationsforschung der Universität Bonn (IPK) der Jahre 1964/65 zurück. Inzwischen hat das IPK seine Schreibkonventionen geändert (vgl. den Forschungsbericht 1966/2 des IPK, Teil II S. 25 ff.); bei uns wurden sie jedoch beibehalten. Eine wechselseitige Umformung ist, mit geringen Abweichungen, möglich.

61) Hier sind nur solche Abweichungen berücksichtigt, die entweder einen Zuwachs oder eine Minderung an Informationen beinhalten.

62) Im ND 64 wurde der Unterschied zunächst noch berücksichtigt, jedoch schon in den Zusätzen zum ND 64 und dann im ND 54 und WELT 64 fallengelassen.

63) Auch System und Form der Transkriptionszeichen wurden vom IPK übernommen, allerdings den differenzierten Ansprüchen unserer Zeitungstexte entsprechend ausgebaut und z.T. geändert. Es wurde dann mit einigen Änderungen auch von der Mannheimer Zentrale übernommen.

64) Abweichend von der Mannheimer Regelung werden Transkriptionen nur durch einfaches Plus-Zeichen geschlossen, das Symbol der Transkriptionsart also nicht wiederholt. Überlappende Transkriptionen, die allerdings sehr selten vorkommen, sind also nach unseren Regeln nicht möglich.

65) Im ND 64 wurden auch eingliedrige Sach-Eigennamen, z.B. Ortsnamen, noch transkribiert. In WELT 64 wurde diese Transkribierung bei eingliedrigen und eindeutigen Sach-Eigennamen jedoch aufgegeben, da sie zweifelsfrei auch aus dem Index zu entnehmen sind, also ein Informationsverlust nicht zu befürchten ist.

66) In WELT 64 sind Buchstabenabkürzungen auch ohne Transkription formal erkennbar, nämlich durch die mehrfache Großschreibung in einem Wort (vgl. Anhang IV, Programmübersicht Nr. 9).

67) Für nicht schreibbare Stellen in der Zeitung (Bilder, Zeichnungen u. dergl.) benutzen wir das Transkriptionszeichen +X...+, wobei die Transkription mit einer Stichwortangabe dessen, was an dieser Stelle in der Zeitung steht, gefüllt wird. Für nicht zu schreibende Stellen (Zahlenkolonnen u. dergl.) wird in entsprechender Weise die Transkription +Y verwandt.

68) Die Informationskarte mit ihren Erweiterungen wurde, aufgrund von Anregungen des IPK, im wesentlichen in der Außenstelle neu erarbeitet.

69) Falls ja, kann die Artikelnummer, unter der der fehlende Teil nachträglich aufgenommen wird, auf einer besonderen Karte vermerkt werden (siehe unten).

70) Unsere Gliederung in 5 Hauptsachgebiete mit je 20 bis 28 Unter-Sachgebieten ist rein pragmatisch; sie erhebt keinen Anspruch auf systematische Folgerichtigkeit und Vollständigkeit. Sie entstand im übrigen in Anlehnung an das von Heinz Vater entwickelte und im Lüneburger Forschungsbericht des Jahres 1965 vorgelegte System von Sachgebieten, wurde allerdings erheblich abgeändert.

71) Vgl. oben 2.1., Abs. 4, Anm. 8.

72) Vgl. oben Punkt 12 (Ergänzungen zu offiziellen Reden). Die Möglichkeiten der Beikarte wurden bisher nur teilweise bei ND 64, die der Zusatzkarte noch gar nicht realisiert.

73) Die Verwendung von Lochstreifen in größeren Mengen ist in Bonn aus überwiegend organisatorischen Gründen leider nur unter großen Schwierigkeiten möglich. Die Datenerfassung im IIM ist ganz überwiegend auf Lochkarten eingestellt. Texterfassung über Streifen hat den Vorteil, daß Streifen erheblich einfacher aufzubewahren sind als Lochkarten, die teure und platzverbrauchende Schränke benötigen; demgegenüber bietet die Verwendung von Karten vor allem beim Korrigieren große Vorteile.

74) Die beiden Lochkarten-Schreibgeräte sind gemietet.

75) Zu Anfang wurde neben dem Schreiblocher ein Lochkartenprüfer IBM 056 verwendet. Das Kartenprüfen mit dem genannten Gerät erwies sich jedoch als unzweckmäßig, da mehr (scheinbare) Fehler angezeigt als wirkliche beseitigt wurden. Da eine Prüfung "von Hand" wegen unserer Transkriptionen ohnehin nötig ist, wurde Ende 1965 auf den Kartenprüfer zugunsten eines zweiten Schreiblochers verzichtet.

76) Auf der Mannheimer Herbsttagung 1965 des IDS wurden, allerdings auf der Grundlage begrenzten Textmaterials, folgende spezielle Register aus kombinierten Abfragen vorgelegt:
1. Alphab. Index der Artikel zum Sachgebiet "Wirtschaft" (=Abfrage nur nach Inform.-Karte)
2. Verzeichnis der Abkürzungen in Reden W. Ulbrichts (=2 verschiedene Abfragen aus Inform.-Karte und Abfrage nach Abkürzungs-Transkription),
3. Verzeichnis der Personennamen (mit Titeln) aus Artikeln über die Bundesrepublik (=Abfrage nach Beikarte und Abfrage nach Personennamen-Transkription).

Auf der Jahrestagung März 1968 wurden u.a. als Probeläufe vorgelegt:
1. Spezial-Index der Punkt-Abkürzungen aus ND 64,
2. Kontext-Register der Initial-Abkürzungen aus WELT 64.

77) Im Index zum ND 64 fanden sich noch ca. 40 Fehler.

78) Zunächst nur im ND 64.

79) Näheres s. oben 4.3., Anm. 17 und Programmübersicht Nr. 14 u. 15.

80) Auf der Jahrestagung März 1968 wurden zwei Probeläufe dieser Art vorgelegt, von denen der eine die Suche nach "Wörtern" (auch nach Wortteilen, sofern die gesuchte Zeichenkombination im Innern zu sammengesetzter Wörter steht), der andere die Suche nach "Wortendungen" (in diesem Fall Substantiv-Suffixen) demonstrierte. Das Verfahren der Suche nach Endungen befriedigt noch nicht, da "Wortendung" zur Zeit noch definiert ist als Zeichenfolge, der mindestens zwei Konsonanten und ein Vokal vorausgehen und ein Leerzeichen folgen müssen.

81) Im Normalfall werden genaues Erscheinungsdatum und Seite des Artikels angegeben. Vgl. Programmübersicht, Nr. 7a.

82) Ungenauigkeit für die Werte von S_{eJ} und S_{aJ} : $\pm$ 20 Seiten. Diese mögliche Abweichung wirkt sich in der Berechnung der Seitenverteilung in der Modellmenge nicht mehr aus. - Die Abweichung gilt auch entsprechend für die Ermittlung der Werte für WELT 64 und ND 54.

83) Entsprechend der unter 4.5.4.2. vorgeschlagenen vereinfachenden Möglichkeit wird hier sowohl für die Berechnung von F als auch von f nicht der (ermittelte) Wert von S_{aJ} (=2125), sondern der (tatsächlich ausgezählte) Wert aus 5 Monaten (=912) zugrundegelegt. Genauer müßte also statt S_{aJ} ein anderer Begriff (etwa S_{aT} wenn T die ausgezählte Teilmenge ist) gesetzt werden. - Entsprechendes gilt für die Modellmengenberechnung der anderen beiden Jahrgänge.

84) Einschließlich unnumerierter Seiten, jedoch ausschließlich aller Beilagen (vgl. oben 8.1.b und 4.5.3.1.a, Anm. 28).

85) Zahl der Erscheinungsmengen und der Ausgaben stimmen im ND 54 (wie schon in WELT 64) überein, da zwar keine Montagsausgabe, jedoch eine Sonntagsausgabe erschienen ist, die mitberücksichtigt wurde.

III. Über die Signifikanz von Auswahltexten
Untersuchungen auf der Grundlage von Zeitungstexten

von Günther Billmeier

Übersicht

Grundlage jeder linguistischen Untersuchung ist eine Textmenge, deren Umfang von der Fragestellung und den finanziellen und technischen Gegebenheiten abhängt. Allgemein lassen sich zwei Methoden unterscheiden:

1. Für die Untersuchung wird der komplette Grundtext aufgenommen (z.B. bei den Kant-, Goethe- oder Büchnerkonkordanzen).
2. Aus dem Gesamttext werden Teile ausgewählt, die die Untersuchungsgrundlage bilden.

Die zweite Art der Textaufnahme wird stets dann praktiziert werden müssen, wenn der Grundtext Größenordnungen zeigt, die eine Verarbeitung nicht mehr zulassen. Die Gültigkeit der am Teiltext gewonnenen Ergebnisse für den Gesamttext muß dann begründet werden; eine Froderung, der in manchen Forschungsarbeiten mit ungenauen Formulierungen der Art "die beobachtete Textmenge scheint ausreichend groß zu sein" genügt wird. Tatsächlich existieren, soweit dem Verfasser bekannt, keine Verfahren, mit deren Hilfe die Abhängigkeit von Grund- (Gesamt-) und Teiltext bestimmt werden kann. [1)]

In der vorliegenden Arbeit soll der Versuch unternommen werden, das Verhältnis eines Teiltextes zum Gesamttext auszudrücken und diese Beziehung als Maß für die Genauigkeit des Teiltextes - bezogen auf den Gesamttext - oder als Grad der Abhängigkeit des Teiltextes zu definieren. Es wird angenommen, daß die Texte nur durch ihre Länge - also die Anzahl der Wörter (Token) - bestimmt sind, so daß es darum gehen muß, eine Bezugsgröße zu finden, deren Verhalten für beliebige Textmengen vorausgesagt werden kann. Textverhältnisse werden dabei durch die Veränderung der relativen Häufigkeiten dargestellt; ein Verfahren, das auf der Annahme basiert, daß die absoluten oder relativen Häufigkeiten der Wortformen (Types) ein Textkriterium darstellen und Textvergleiche

zulassen. Aufgabe war es, festzustellen, ob die vom Institut für deutsche Sprache, Außenstelle Bonn, nach den dort entwickelten Prinzipien der (qualitativen) Aufnahme von Zeitungstexten verkartete Textmenge für repräsentative Aussagen genügt. Alle Ergebnisse können deshalb nur für das untersuchte Material gültig sein. Mein besonderer Dank gilt Herrn Professor Dr. Hugo Moser und dem Institut für deutsche Sprache, die es mir ermöglichten, die ursprünglich nur als Kurzuntersuchung gedachte Arbeit in der vorliegenden Form durchzuführen, sowie Herrn Dr. D. Krallmann für seine Beratung bei den statistischen Problemen.

Die Auswertung der Testdaten erfolgte auf der Großrechenanlage des IIM/IAM der Universität Bonn (System IBM 7090); die verwendeten Programme sind in der Programmiersprache Fortran II geschrieben und vom Verfasser erarbeitet.

Das Verhältnis Teiltext - Gesamttext

Bei der Verkleinerung einer Textmenge geht man von der sicherlich zutreffenden Voraussetzung aus, daß eine Teilmenge genügender Größe die Textverhältnisse (Wortverteilungen, Grammatik, Syntax) des Gesamttextes widerspiegelt und damit verbindliche Aussagen über das Corpus zuläßt.
Die Linguistik bezeichnet einen aus einem Gesamttext ausgewählten Teil als "sample" und unterscheidet - nach Herdan [2] - zwischen

a. der reinen Zufallsauswahl - random sampling -

b. der "Streumethode" - spread sampling - .

Da bei der Beobachtung sprachlicher Kriterien - außer vielleicht bei reinen Wortstatistiken - immer gerade der laufende Text, zumindest über eine bestimmte "Textstrecke" (Phrase, Satz usw.) hin interessiert, kommt als Methode des "sampling" nur die Auswahl bestimmter zusammenhängender Textteile in

Frage. Die Statistik spricht in diesem Fall von einer "Klumpenstichprobe", wobei unter "Stichprobe" hier eine "Teilmenge von Einheiten" verstanden wird, die "aus einer bestimmten Masse ausgewählt wurden, mit dem Ziele, daraus Schlüsse auf die Beschaffenheit dieser Masse zu ziehen". [3]

Die Art des "spread sampling" bringt es mit sich, daß die mathematische Theorie für reine Zufalls-samples hier nicht anwendbar ist:
"Since we necessarily have to sample here by contiguous pieces of writing, which may contain systematically more words of a certain kind, i.e. the words required for developing a given idea or describing a certain event, some sort of bias may be introduced in this way." [4]

Unschwer läßt sich erkennen, daß diese Überbetonung gewisser Erscheinungen im Text, die durch die Art der Auswahl bedingt wird, zu Lasten der seltenen Ereignisse geht. Die naheliegende Lösung, Teil- und Gesamttext durch den Quotienten ihrer Längen in Beziehung zu setzen und das Ergebnis, das ja stets kleiner 1 sein muß, als Wahrscheinlichkeit zu interpretieren, erweist sich damit als unanwendbar. Ohne Zweifel würde dies für Mengen gelten, bei denen eine Vergrößerung linear auf die enthaltenen Teilmengen (hierunter könnte man im sprachlichen Bereich die Wortformen verstehen) wirken würde; für Texte zeigt allein die Betrachtung des Type-Token Verhältnisses, daß diese einfache Quotientenbildung den tatsächlichen Gegebenheiten nicht entspricht.

Zweifellos ist das Vokabular (die Anzahl der Wortformen) eine Funktion der Textlänge, die durch

$$V = T^c$$

angenähert ausgedrückt werden kann, wobei c notwendigerweise mit wachsender Textlänge sinkt, da ein neu hinzukommendes Wort umso eher einer vorhandenen Wortform zugeordnet werden kann, je mehr Wortformen bereits auf-

getreten sind. (c-Werte der untersuchten Textmengen siehe Anhang I) Eine Textvergrößerung verändert demnach die inneren Textverhältnisse nicht linear. Es liegt nahe, für zwei verschieden lange Texte ein Verhältnis der Form

$$V = \left(\frac{T_{gesamt}}{T_{teil}}\right)^{c}$$

anzunehmen und zu versuchen, den Potenzfaktor c für beliebige Texte bestimmbar zu machen. Die Textstatistik tut dies, indem sie verschiedene Texte untersucht und sie "rückblickend" in Beziehung zueinander setzt. Mit einem etwas veränderten Ansatz wollten wir versuchen, die einzelne Wortform in ihrem Verhalten innerhalb eines Textes zu betrachten, sie also gleichsam beim Wachsen des Textes zu seiner Endlänge hin zu beobachten. Wir gingen dabei von der relativen Häufigkeit des Types aus und stellten fest, wie stark diese bei einer beliebigen Textlänge von der der Gesamtlänge abwich. Die "beliebige Textlänge" wurde als Länge des Teiltextes definiert und der "repräsentative" oder signifikante Teiltext folgendermaßen bestimmt:

Ein Teiltext ist repräsentativ für den Gesamttext, wenn ein zu bestimmender Prozentsatz der Worte (1. Setzung) unter einer zu bestimmenden Abweichung (2. Setzung) liegt und die einzelne Wortform die Abweichungsgrenze bei Textvergrößerung nicht mehr überschreitet.

Grundgedanke dieser Arbeitshypothese war die folgende Überlegung: In einem Text besitzt ein Type eine bestimmte Häufigkeit F. Wird derselbe Type in einem kürzeren Text, der aus dem langen ausgewählt wurde, untersucht und soll der verkürzte Text ein getreues, verkleinertes Abbild des Gesamttextes sein, so müßte der Type hier die Häufigkeit $\frac{F}{n}$ ($n = \frac{\text{Länge Gesamttext}}{\text{Länge Teiltext}}$) besitzen. Unschwer ist zu erkennen, daß diese Annahme nur im Bereich der großen und mittleren Häufigkeiten zutreffen kann, und daß Schwankungen zwischen $\frac{F}{n}$ und der tatsächlichen Häufigkeit im Teiltext auftreten werden.

Die Schwankung, als Abweichung bezeichnet, sinkt, je mehr der Teiltext sich in seiner Länge dem Gesamttext nähert und wird beim Gesamttext gleich Null.

Betrachtet man nun in einem Teiltext alle im Gesamttext vorhandenen Types (Wortformen, die erst bei Textvergrößerung auftreten, haben hier dann den höchstmöglichen Abweichungswert) daraufhin, ob sie unter einer beliebigen Grenze der Abweichung liegen, so erhält man in der Summe der Häufigkeiten dieser Types den Prozentsatz der Worte des Teiltextes, die die Bedingung erfüllen. Im idealen, signifikanten Teiltext wären es 100 Prozent der Worte bei einer Abweichung von Null.

Die Abweichungsgrenze und der geforderte Prozentsatz sollten empirisch ermittelt werden.

Der erste Abschnitt des vorliegenden Berichts behandelt die theoretischen Grundlagen der Untersuchung, der zweite schildert die praktische Durchführung einschließlich der Programmierung; die Ergebnisse und die damit durchgeführte Berechnung der Konstanten enthält Teil 3.

1. Abschnitt

Die theoretischen Grundlagen der Untersuchung

Die Abweichung (A)

1. Begriffsbestimmung und Berechnung

Jede Wortform besitzt in einem Text eine bestimmte relative Häufigkeit, die sich durch den Quotienten von absoluter Häufigkeit und der Anzahl der Gesamtworte (Textlänge) bestimmen läßt.

Betrachtet man die Werte der relativen Häufigkeit in verschieden langen Teiltexten und im Gesamttext, so ergeben sich Differenzen, die das Verhältnis eines Types im Teiltext zum Gesamttext ausdrücken. Diese Differenzen bezeichnen wir als Abweichung.

T_n sei eine Textmenge mit N Token. Aus ihr werden n Teile zu je $\frac{N}{n}$ Token gebildet, die wiederum zu neuen Teiltexten zusammengefaßt werden, wobei für die Länge gilt:

$$T_1 = \frac{N}{n} \quad \text{Token}$$

$$T_2 = 2 \cdot \frac{N}{n} \quad \text{Token}$$

$$T_k = k \cdot \frac{N}{n} \quad \text{Token} = K \text{ Token}$$

$$T_n = n \cdot \frac{N}{n} \quad = N \text{ Token}$$

In Analogie zum untersuchten Text wird in der Arbeit die Textmenge $\frac{N}{n}$ als Seite bezeichnet. k und n sind daher in den Formeln durch $k \cdot \frac{N}{n}$ bzw. $n \cdot \frac{N}{n}$ ersetzbar.

$F(w,T_i)$ sei die absolute Häufigkeit eines Types w im Textstück

$$T_i = i \cdot \frac{N}{n} \quad \text{Token}$$

$f(w,T_i)$ die relative Häufigkeit desselben Types bei T_i

Die Abweichung A' eines Types w im Textstück T_k ($k \leq n$) ist dann die Differenz aus der relativen Häufigkeit bei T_n und der relativen Häufigkeit bei T_k .

$$A'(w,T_k) = \left| f(w,T_n) - f(w,T_k) \right| \quad (1)^{5)}$$

(1) bedingt die Abhängigkeit des Wertes von $f(w,T_n)$. Je größer die relative Häufigkeit eines Types, umso größer ist auch seine Abweichung. Zur Überführung in eine davon unabhängige Form multipliziert man A mit dem Reziprokwert von $f(w,T_n)$ und - zur Erzielung praktikablerer Werte - 100 und erhält A in Prozenten von $f(w,T_n)$.

$$A(w,T_k) = 100 \cdot \left| \frac{f(w,T_n) - f(w,T_k)}{f(w,T_n)} \right| \quad (1a)$$

Setzt man

$$f(w,T_i) = \frac{F(w,T_i) \cdot 100}{T_i} = \frac{F(w,T_i) \cdot 100}{i \cdot \frac{N}{n}}$$

in (1a), so ergibt sich

$$A(w,T_k) = 100 \cdot \left| \left(1 - \frac{F(w,T_k) \cdot n}{F(w,T_n) \cdot k} \right) \right| \quad (1b)$$

A ist damit nicht von der relativen oder absoluten Häufigkeit eines Types im Gesamttext abhängig. Dies gilt jedoch nur solange, als $F(w,T_n)$ mindestens gleich n ist. Wird $F(w,T_n)$ kleiner als die Anzahl der Textstücke, so gewinnt $\frac{n}{k}$ an Bedeutung, was bei den Hapaxlegomena zu einer Verkürzung von (1b) auf

$$A(w,T_k) = 100 \qquad \text{(vor dem Auftreten)}$$

und

$$A(w,T_k) = 100 \cdot (1 - \frac{n}{k}) \qquad \text{(nach dem Auftreten)}$$

führt. Ähnlich sind die Verhältnisse bei den 2, 3 .... mal auftretenden Worten.

Abbildung 1 zeigt das Verhältnis zwischen dem erwarteten Vorkommen ($= \frac{F(w,T_n)}{n} \cdot k$), dem tatsächlichen Vorkommen ($= F(w,T_k)$) und der Abweichung (= A).[6)]

Exkurs

Die Abweichung der einzelnen Wortform als Mittel zur Erkennung textsignifikanter Wörter

Geht man von der Voraussetzung aus, daß ein Teiltext das verkleinerte Abbild des Gesamttextes darstellt, so müßte die relative Häufigkeit eines Types im Teiltext gleich der "Schlußhäufigkeit" sein, die Wahrscheinlichkeit für das Auftreten eines zu diesem Type gehörenden Token also gleich bleiben. In einer Darstellung Textmenge/kumulierte Häufigkeit ergibt sich dann eine Gerade, deren Steigungswinkel von der Höhe der "Schlußhäufigkeit" abhängt. (Siehe auch Abb. 1)

Taucht ein Type in einem Teiltext plötzlich mit stark erhöhter oder verminderter Häufigkeit auf (i.e. Kurve 2 der Abb. 1 weicht stark von 1 ab, 3 fällt oder steigt stark[7]), so weist dies auf einen besonders häufigen (oder seltenen) Gebrauch des Wortes in eben diesem Teiltext hin. Handelt es sich um ein Substantiv, Adjektiv oder Verb, liegt der Schluß vom Wort zum Inhalt nahe.

Das Gesagte soll - leider verbietet die Begrenzung des Themas, näher auf diese Erscheinung einzugehen - durch ein Beispiel erläutert werden.

Eine Untersuchung der A-Werte ergab für den Type "ökonomischen" einen starken Sprung vom 3. zum 4. Teiltext ($H = 852$, $T_3 = 9\,000$, $T_4 = 12\,000$ Token). Eine Prüfung der übrigen Wortformen auf den größten Sprung ergab, daß "Leitung", "Planung", "System(s)", ökonomische" und "Volkswirtschaft" beim selben Teiltext die größten Unterschiede ihrer A-Werte aufweisen.[8]

Ein Suchprogramm lieferte anschließend alle Belege der genannten Wörter im Gesamttext, und man fand, daß im 4. Teiltext (= Seite 4 des Textmaterials) der Terminus "das neue ökonomische System der Planung und Leitung der Volkswirtschaft" zum ersten Mal eingeführt und von Ulbricht in seinem Referat vor dem V. ZK-Plenum (im Text ist nur ein Auszug des Referats vorhanden) 30 mal gebraucht wurde (3 Nom., 6 Gen., 1 Dat.; verkürzt zu "neues ökonomisches System": 8 Nom., 8 Gen., 2 Dat., 2 Akk.).

2. Bedingung und Erfüllung

Die Bedingung ist eine gesetzte Abweichung, die eine Wortform unterschreiten muß - und in keinem längeren Teiltext mehr überschreiten darf - , um als schwankungsfrei zu gelten.

$$B \leqq 100 \cdot \left| \left(1 - \frac{F(w,T_k)\cdot n}{F(w,T_n)\cdot k} \right) \right|$$

Die Wortform ist schwankungsfrei, wenn :

$$(n-k)\cdot B \leqq \sum_{j=k}^{n} 100 \cdot \left| \left(1 - \frac{F(w,T_j)\cdot n}{F(w,T_n)\cdot j} \right) \right|$$

(T_n fest)

Für die empirische Ermittlung von B kommen die Types mit großen Häufigkeiten in Frage, da diese sich mit größter Wahrscheinlichkeit "ideal" verhalten. Die Betrachtung der 30 häufigsten Types des Untersuchungstextes ergab, daß - von wenigen Ausnahmen bei sehr geringer Textmenge abgesehen - die Schwankungen mit einem B-Wert von 25 (Prozent) hinreichend erfaßt werden .

Es muß darauf hingewiesen werden, daß dieser Wert als erster Ansatz gedacht war, und daß er für die häufigsten Types als zu hoch erscheinen muß. Bedeutet er doch, daß z.B. "die" mit einer relativen Häufigkeit zwischen 3 und 6 % auftreten könnte. Da der weitaus überwiegende Teil der Types geringere Häufigkeiten aufweist und die häufigsten Wortformen für die Untersuchung nicht von Bedeutung sind (ihr "ideales" Verhalten wird ja vorausgesetzt), muß diese Schwankungsbreite vorgegeben werden, um auch die selteneren Wortformen - mit ihren naturgemäß höheren Schwankungen - zu berücksichtigen.

Betrachtet man die Höhe der Abweichung in Teiltexten wachsender Länge (zum Gesamttext hin), so sieht man, daß die Abweichung umso kleiner wird, je mehr der Teiltext sich dem Gesamttext nähert. Jeder Type muß also - spätestens beim Gesamttext - die Abweichung 0

erreichen, bzw. einen Wert, der unter der gesetzten Grenze liegt. Untersucht man in einem beliebigen Text T_k ($k \leq n$) alle Types daraufhin, ob sie unter B liegen, so erhält man die Anzahl der Wortformen, die die Bedingung erfüllen.

Der Prozentsatz, den diese Types im gerade untersuchten Text ausmachen, nennen wir Erfüllung.

Die Erfüllung wird in Prozenten der Textmenge gemessen; wählt man B hinreichend groß, ist sie in jedem Teiltext gleich 100. In einem P/T- (Erfüllung/Textmenge) - Diagramm ergibt sich damit eine Kurve, die durch die linke untere (0/0) und rechte obere ($100/T_n$) Ecke führen muß und deren Steigung stets positiv ist.

Nach der Definition für die Länge eines Teiltextes (Seite 132) ist $T_k = T_{k-1} + \frac{N}{n}$; jeder größere Teiltext enthält $\frac{N}{n}$ Token mehr als der vorhergehende. Die bereits vorhandenen Types wachsen dadurch in ihrer absoluten Häufigkeit und neue Types treten hinzu. Daraus folgt, daß die Erfüllung erst bei $T_k = T_n$ gleich 100 werden kann, da die im letzten Textteil neu auftretenden Wortformen mit ihren A-Werten über B liegen können.

Nach der auf Seite 134 angegebenen Formel für das Verhalten der Hapaxlegomena, bzw. der seltenen Wortformen, kann im letzten Drittel der P/T-Verteilung ein Sprung angenommen werden. Die seltenen Wortformen sind hier mit größter Wahrscheinlichkeit bereits voll aufgetreten und sinken damit an einer bestimmten Stelle des Textes in ihrem A-Wert unter B.
Hat ein Type in einem Text die absolute Häufigkeit 5, so ist die Wahrscheinlichkeit für das Auftreten eines zu diesem Type gehörenden Token im Textstück T_k mit der Länge $\frac{T_n}{5}$ gleich 1, d.h. in 20 % des Gesamttextes wird wahrscheinlich ein Token enthalten sein. Betrachtet man nun

einen Text, der über 80 % des Gesamttextes ausmacht, so kann in ihm die Endhäufigkeit des Types bereits erreicht sein, oder: spätestens in den letzten 20 % des Textes muß der Type seine volle Häufigkeit erhalten. Die Abweichung wird damit:

$$A = 100 \cdot (1 - \frac{n}{k})$$

Bei gesetzter Bedingung gilt demnach:

$$B \leqq 100 \cdot (1-\frac{n}{k})$$

oder - siehe Anmerkung 5) -

$$B = 100 \cdot (\frac{n}{k} - 1)$$

k ist dabei diejenige Textmenge, bei der der Sprung auftritt. Aufgelöst ergibt sich:

$$k = \frac{100.n}{100+B}$$

Als "selten" wird nun eine Wortform definiert, die bei der Textlänge, bei der der Sprung erfolgt, bereits mit ihrer Endhäufigkeit aufgetreten sein kann. Die Häufigkeit der seltenen Wortform berechnet sich dann:

$$F_{slt.\ Wf} \leqq \frac{k}{1-k}$$

Ersetzt man n durch 1, d.h. wird der Gesamttext eine vom Wert unabhängige Bezugsgröße, so ist

$$F_{slt.\ Wf.} \leqq \frac{100}{B}$$

(Liegt der Absprungpunkt k z.B. bei 75 % der Textmenge - k = 0,75 -, so können alle Types, die im Gesamttext nicht öfter als drei Mal auftreten, ihre Endhäufigkeit erreicht haben.)

Absprungpunkt und Häufigkeit der seltenen Wortform sind damit als von der Textlänge unabhängige Größen definiert. Je größer die gewählte Abweichungsgrenze B ist, desto kleiner wird die absolute Häufigkeit eines "selten" genannten Types; sinkendes B läßt auch die mittleren und großen Häufigkeiten zu "seltenen" Wortformen werden, bis bei B = 0 jeder Type als "selten" anzusprechen ist.

Die Sprunghöhe muß gleich der relativen Häufigkeit aller Types sein, deren absolute Häufigkeit im Gesamttext zwischen 1 und $\frac{100}{B}$ liegt. Bei bekanntem Endzustand eines Textes (vorliegendes Häufigkeitsregister) kann dieser Wert leicht ermittelt werden; für unbekannte Texte soll versucht werden, das Verhalten der seltenen Wortformen zu bestimmen.

3. Die Kennlinie

Bei gegebener Abweichungsgrenze B liegt der Sprung der P/T- Verteilung immer an der gleichen Stelle. Er unterscheidet sich bei verschiedenen Texten lediglich durch die Höhe des Absprungpunktes, die ihrerseits durch die Höhe des Sprungs, also den Anteil der seltenen Wortformen bestimmt wird.

Der Verlauf einer P/T-Verteilung bei gegebenem B soll Kennlinie genannt werden. Die Kennlinie gibt das Verhalten der Wortformen bei wachsender Textlänge in Abhängigkeit von der Abweichung wieder; sie ist - wie der Anteil der seltenen Wortformen - ein Textparameter. Sie läßt Aussagen über das Vokabular eines Textes und seine Ausnutzung zu (je besser und gleichmäßiger ein vorhandener Wortschatz ausgenutzt wird, umso kleiner die Sprunghöhe; je größer das Vokabular, umso größer die Sprunghöhe) und gibt Aus-

kunft darüber, bei welchem Teil des Textes sich welcher Prozentsatz der Token dem Endzustand angeglichen hat. Sie ist vom individuellen Text abhängig und verändert sich mit der Textlänge.

4. Das Verhalten der seltenen Wortformen

Die seltenen Wortformen beeinflussen die P/T-Verteilung im letzten Drittel in direkter Abhängigkeit von k (Damit wieder von B.) Je kleiner B, desto größer die absolute Häufigkeit eines als "selten" bezeichneten Types. Bei B = 0 ist jeder Type als selten definiert, da keiner vor dem Erreichen der Endhäufigkeit völlig schwankungsfrei ist.

Notwendigerweise sinkt der Anteil der seltenen Wortformen bei wachsender Textlänge; je größer die Anzahl der bereits vorhandenen Types ist, umso größer ist auch die Wahrscheinlichkeit, daß ein neu auftretendes Wort einem bereits vorhandenen Type zugeordnet werden kann und nicht eine neue Wortform darstellt.

Es kann eine Näherungsgleichung angesetzt werden:

$$S = T^{a}$$

(Der Anteil der seltenen Wortformen ist eine Funktion der Textlänge.)

Bei diesem Ansatz ergibt sich der Anteil der seltenen Wortformen als absolute Zahl. Da bei dieser Untersuchung der Anteil der seltenen Wortformen in Prozenten der Gesamtwortmenge interessiert, soll gesetzt werden:

$$S = \frac{M}{T^{a}} \qquad (2)$$

Zugleich ist S abhängig von B, so daß weiter gilt:

$$M = f(B)$$

Der Anteil der nicht seltenen Wortformen berechnet sich dann durch:

$$NS = 100 - \frac{M}{T^a}$$

Hierbei geht man davon aus, daß nach dem Sprung sofort alle Types der seltenen Wortformen unter B liegen; es treten aber auch dann noch Wortformen auf, die über B liegen - textsignifikante Wörter der letzten Textteile - , so daß mit

$$NS = 95 - \frac{M}{T^a} \qquad (3)$$

genauere Ergebnisse möglich werden.

(2) und (3) ermöglichen die Berechnung des Schnittpunktes:

$$T_{Schnitt} = \sqrt[a]{\frac{M}{47,5}}$$

Bei $T_{Schnitt}$ ist der Anteil der seltenen Wortformen gleich dem der nicht seltenen. Eine weitere Textverkleinerung läßt keine Aussagen über Textverhältnisse mehr zu.

Das bestimmte Integral

$$\int_{T_1}^{T_2} \frac{M dT}{T^a}$$

ist die Fläche, die ein Textstück der Länge T_2 mit einem Text der Länge T_1 zwischen der x-Achse und der Kurve der seltenen Wortformen einnimmt. (Siehe Abb. 2) Setzt man T_1 als die untere Grenze der Textmenge, die eine Aussage zuläßt - also $T_{Schnitt}$ - , so bildet die eingeschlossene

"Textfläche" einen weiteren Textparameter.

Der Anteil der seltenen Wortformen wurde im Vorhergehenden als wichtiges Kriterium eines Textes bestimmt. Man kann die "Textfläche" daher als Aussage über den Text auffassen und "Teiltextflächen" zur "Gesamtfläche" in Bezug setzen.

Das Verhältnis

$$\frac{\text{vom Teiltext gebildete Kurvenfläche}}{\text{vom Gesamttext gebildete Kurvenfläche}}$$

oder

$$\frac{\int_{T_1}^{T_k} \frac{M}{T_k^a}\, dT_k}{\int_{T_1}^{T_n} \frac{M}{T_n^a}\, dT_n}$$

soll als "V-Quotient eingeführt werden.

Aufgelöst ergibt sich:

$$V = \frac{T_k^{-a+1} - T_1^{-a+1}}{T_n^{-a+1} - T_1^{-a+1}}$$

Gilt

$$T_1 \ll T_k$$

so ist

$$V = \left(\frac{T_k}{T_n}\right)^{-a+1}$$

Teil- und Gesamttext werden nun nicht mehr nur durch ihre Längen in Beziehung gesetzt; ihr Verhältnis beruht jetzt auf der Betrachtung der seltenen Wortformen, durch die die inneren Texterscheinungen beim "Wachstum" berücksichtigt werden. Im Grunde bedeutet dies eine Wiederholung der Theorie, das Vokabular sei eine Funktion der Textlänge, da die Veränderungen in der Größe des Vokabulars auf den seltenen Wortformen beruhen. Im vorliegenden Ansatz ist allerdings versucht worden, eine einzige Erscheinung herauszugreifen und ihre Bedeutung für das Verhältnis der Texte zu begründen.

V kann als Wahrscheinlichkeit interpretiert werden, mit der Erscheinungen des Gesamttextes im Teiltext berücksichtigt sind. Bei gegebenem B-Quotienten errechnet sich die Länge des Teiltextes T_k als

$$T_k = \sqrt[-a+1]{V \cdot T_n^{-a+1} + T_1^{-a+1} (1-V)}$$

angenähert:

$$T_k = T_n \sqrt[-a+1]{V \pm \frac{T_1}{2}}$$

2. Abschnitt

Praktische Durchführung

1. Das untersuchte Textmaterial

Aus dem Jahrgang 1964 des "Neuen Deutschland", Berliner Ausgabe, wurden 18 Zeitungsseiten des Monats Februar nach den am Institut für deutsche Sprache entwickelten Prinzipien der Textaufnahme ausgewählt. Bei insgesamt 24 Ausgaben dieser Zeitung (ohne Sonntage) mit 184 Seiten im Februar entspricht die Auswahl annähernd 10 % der Gesamtmenge (genau 9,78 %).
Diese Textmenge wurde als Ausgangsbasis der Untersuchung, als Gesamttext, angesetzt.

Der so zusammengestellte Text ergibt 60 139 Einzelworte, wobei die Verteilung über die einzelnen Seiten - bedingt durch die Stellung der ausgewählten Seite in der Ausgabe, Schlagzeilen, Bilder, Anzeigen usw. - starken Schwankungen unterworfen ist.

Seite i. d. Ztg.	Datum	Inhalt	Anzahl der Worte
2	1.2.	Politik, Kommentar, 1 Zchng., 1 Faks.	3635
3	1.2.	Aktuelles zur Wirtschaft, 2 Bilder	3143
1	5.2.	Titelseite, 4 Bilder	2007
4	5.2.	Referat Ulbricht auf 5. ZK-Tagung	4589
2	10.2.	Politik, Kommentar, 1 Zchng., 1 Bild	3715
3	10.2.	Diskussion 5. ZK-Plenum	4170

5	13.2.	Kulturberichte, Bericht Polibüro an ZK-Plenum	3419
6	13.2.	Anzeigen	1628
8	13.1.	Sport, Wirtschaft, Wetterberichte, Spielplan, 2 Bilder 2 Zchngn.	2673
2	17.2.	Politik	3993
7	17.2.	Sport, 2 kleine Zchngn.	3841
4	21.2.	Kultur, 1 Bild	3428
6	21.2.	Amerika-Bericht, 3 Bilder	2828
4	25.2	Kultur, 2 Bilder	3456
5	25.2.	Politik, 1 Bild	3768
1	29.2.	Titelseite, 1 Bild	1923
7	29.2.	Referat Ulbricht auf dem 8. deutschen Bauernkongreß	4465
8	29.2.	Sport, Referat Ulbricht Fortsetzung, 1 Bild	3458
18			60 139

Zur Vereinfachung der Auswertung nahm man eine neue Einteilung in 20 Seiten (Textstücke) zu je 3 000 Wörtern vor, so daß 60 000 Token zur Bearbeitung zur Verfügung standen.

2. Maschinelle Verarbeitung

Mit Hilfe des Programmes TRENN erstellte man einzelne Wortrecords normierter Größe aus dem laufenden Text, wobei Zahlen, Satzzeichen und vom Bearbeiter eingefügte Sonderzeichen oder Worte getilgt wurden.

Ein Programm INDIZ faßte jeweils 3 000 Wortrecords zu einem Block zusammen, sortierte ihn und gab ihn an INDEX weiter, das daraus ein Wörterbuch - nur Wort, absolute Häufigkeit und Nummer des Blocks (1 - 20) anfertigte. In einem letzten Programm FILSOR wurden die 20 Wörterbücher nach Wort und Blocknummer sortiert, so daß das fertige Band für jeden Type - sofern er in einem der Blöcke auftauchte - einen, oder, entsprechend seinem Auftreten, bis zu 20 Records enthielt.

Das fertige Band hatte folgendes Aussehen :

LOESBAR0000000000000000000000000001000002
LOESEN0000000000000000000000000003000004
LOESEN0000000000000000000000000003000007
LOESEN0000000000000000000000000001000009
LOESEN0000000000000000000000000002000018
LOESEN0000000000000000000000000001000019
LOESEN0000000000000000000000000002000020
LOEST 0000000000000000000000000003000004
LOESUNG000000000000000000*00005000003
LOESUNG000000000000000000*00002000004
LOESUNG000000000000000000*00003000007
LOGAN-*BEZIRK00000000000000*00002000014

Die ersten Stellen enthalten das Textwort, (Type), die folgende Zahl gibt die absolute Häufigkeit im entsprechenden Block an; es folgt die Blocknummer. Der Stern zeigt Großschreibung an, die Nullen entstehen durch maschinelle Verarbeitung. Seite 148 zeigt das Flußdiagramm für die Aufbereitung des Textes.

Das Auswertprogramm faßt die zu einem Type gehörenden Records zusammen - fehlende Records (= Type tritt in einem Block nicht auf) werden durch 0 ersetzt - und fertigt eine Liste (siehe Seite149) an. In dieser Liste prüft es, von welchem Textstück an der Type eine gesetzte Abweichung nicht mehr überschreitet. Den Wert der relativen Häufigkeit, den der Type in diesem Textstück hat, addiert es in den entsprechenden Platz einer Matrize.

Gleichzeitig ermittelt es das erste Auftreten eines Types und verfügt damit am Ende des Programmlaufes über die Zahl der neuen Wortformen pro Block.

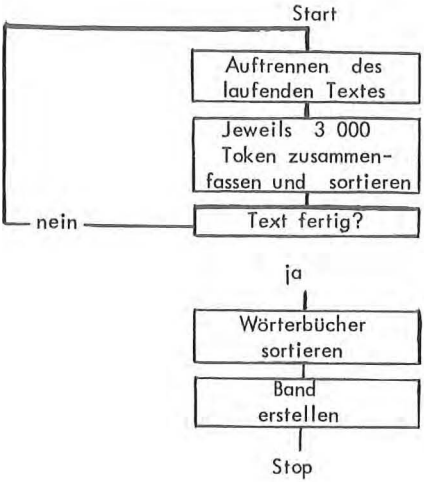
Sollte z.B. in einem Lauf die Abweichungsgrenze B = 25 beachtet werden, so würde "AB" - siehe Seite 149 - ab Seite 13 die Bedingung erfüllen. Die relative Häufigkeit von 0,058974 würde damit in der Matrize bei Seite 13 zuaddiert, wo bereits Werte von früher geprüften Types summiert sein könnten. Selbstverständlich wird die relative Häufigkeit auch bei allen folgenden Seiten (14 - 20) zuaddiert, da ja "AB" auch hier die Bedingung erfüllt.

Seite.....	12	13	14	15	16 .....
	21,3542	24,3648	28,4728	32,1234	39,8342

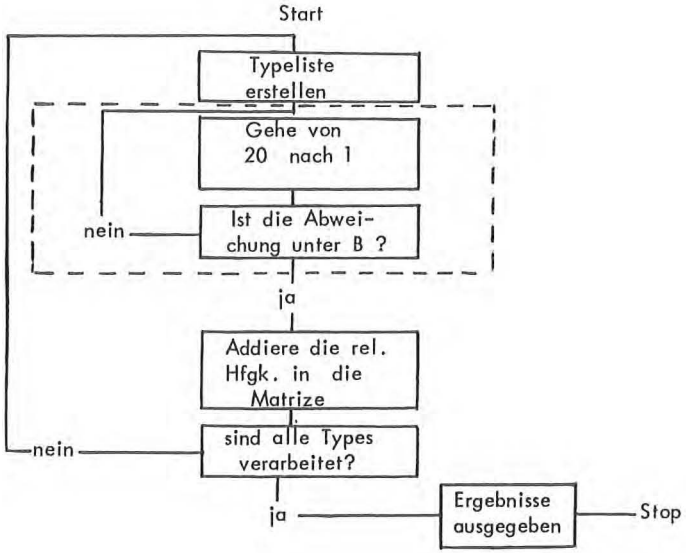
+ Wert "AB" = 0,058974 ab Seite 13

Seite	12	13	14	15	16 .....
	21,3542	24,4238	28,5318	32,1824	39,8932

Matrizenaddition



Aufbereitung des Textes



Auswertung

Maschinell erstellte Liste für den Type "ab"

AB

SEITE	VORKOMMEN 1	KUM. VORK. 2	WORTE/ SEITE 3	GESAMT- WORTE 4	REL.VK. (BEZ. AUF 2 UND 4)	ABWEICHUNG V. ENDWERT
1	1	1	3000	3000	0.033333	31.0345
2	1	2	3000	6000	0.033333	31.0345
3	4	6	3000	9000	0.066667	37.9310
4	0	6	3000	12000	0.050000	3.4483
5	0	6	3000	15000	0.040000	17.2412
6	2	8	3000	18000	0.044444	8.0460
7	0	8	3000	21000	0.038095	21.1823
8	1	9	3000	24000	0.037500	22.4138
9	6	15	3000	27000	0.055556	14.9425
10	3	18	3000	30000	0.060000	24.1379
11	2	20	3000	33000	0.060606	25.3918
12	2	22	3000	36000	0.061111	26.4368
13	1	23	3000	39000	0.058974	22.0159
14	0	23	3000	42000	0.054762	13.3005
15	2	25	3000	45000	0.055556	14.9425
16	3	28	3000	48000	0.058333	20.6897
17	0	28	3000	51000	0.054902	13.5903
18	0	28	3000	54000	0.051852	7.2797
19	0	28	3000	57000	0.049123	1.6334
20	1	29	3000	60000	0.048333	0.

3. Abschnitt

Die Ergebnisse und ihre Auswertung

1. Gesamttext und Teiltexte

Das untersuchte Textmaterial bestand aus 60 000 Token, die in 20 Teile zu je 3 000 Token aufgeteilt wurden.

$$T_n = 60\ 000 = N$$

$$n = 20$$

$$T_1 = 3\ 000$$

$$T_2 = T_1 + \frac{N}{n} = 3\ 000 + \frac{60\ 000}{20} = 6\ 000$$

Der 60 000-Token-Text wurde als Gesamttext aufgefaßt, dem drei Teiltexte zu 15 000, 30 000 und 45 000 Token gegenüber gestellt wurden. Da die Ergebnisse des 45 000-Token-Textes stark von den erwarteten abwichen (siehe Abb. 6), wurde außerdem ein aus 45 000 Worten bestehender Text (ebenfalls "Neues Deutschland" 1964, Berliner Ausgabe, aber nicht Monat Februar) zur Untersuchung herangezogen.

2. Wortformen beim Gesamttext

Den 60 000 Token entsprechen 12 832 Types; ihr Auftreten in den einzelnen Textteilen zeigt Abb. 3. Die 30 häufigsten Types machen 31,62 % der Token aus, was weitgehend mit Ergebnissen anderer Untersuchungen übereinstimmt. In Anhang II ist die Reihenfolge der häufigsten Wortformen im Vergleich zu Meier [9] und Eggers [10] wiedergegeben. Die auffälligsten

Unterschiede treten bei den Artikeln "die" und "der" ("der" an erster Stelle) und dem erst an 18. Stelle stehenden "nicht" (sonst um 10) auf; das erste wahrscheinlich textsignifikante Wort "DDR" steht auf Platz 33. Die Belegung der unteren Häufigkeitsklassen ist in Anhang III dargestellt.

3. B-Werte und Kennlinien

Als B-Werte wählten wir 10, 15, 20, 25 und 30 Prozent; bei 30 000 und 60 000 Token wurden alle Werte von B = 5 bis B = 30 durchlaufen, um genaue Aussagen über die Verschiebung des Sprungpunktes und der Sprunghöhe zu erhalten. Abb. 4, A - D zeigen die gemessenen Kennlinien. Die x-Achse gibt die Länge an, auf der y-Achse ist der Prozentsatz der Worte, die unter der Abweichung liegen, ablesbar. Abb. 5 läßt die Verringerung der Sprunghöhe bei wachsender Textlänge erkennen: die Kennlinien der vier Texte sind hier bei B = 25 ineinander gezeichnet. (Die horizontale Verschiebung des Sprungpunktes ist scheinbar, da die x-Achse variiert.)

Bei allen Kennlinien lassen sich vom "Einsatz-" bis zum "Absprungpunkt" Gerade approximieren, die lediglich in ihrer Steigung verschieden sind. Dieser lineare Verlauf unterstützt die Theorie, die Texte würden in erster Linie von den seltenen Wortformen beeinflußt, die sich ja in der Sprunghöhe bemerkbar machen. Der "Einsatzpunkt" hat bei gleichem B bei allen vier Texten etwa den gleichen Wert (26,7 - 24,9 - 24,7 - 23,5). Da die seltenen Wortformen prozentual sinken, der Anteil der "sofort schwankungsfreien" häufigen Types konstant bleibt, könnte sich die leichte Abnahme im Einsatzwert durch die Erweiterung der mittleren Häufigkeitsklassen, in denen sich die meisten textsignifikanten Wörter befinden (die ja den stärksten Schwankungen unterworfen sind), erklären.

4. Erwartete und tatsächliche Ergebnisse

Aus Raumgründen soll der Vergleich zwischen den tatsächlichen und errechneten Ergebnissen auf den 60 000-Token-Text beschränkt werden.

a. Absprungpunkt

B-Wert	errechneter Absprung jeweils umgerechnet auf Seite	gemessener Absprung
10	18,18	18
15	17,39	17
20	16,70	16
25	16	15
30	15,38	15

Der Absprungpunkt wurde nach der Formel

$$k = \frac{100}{100+B} \cdot n$$

errechnet, so daß die Textlänge in Token beim Absprung gleich

$$T_k = \frac{100}{100+B} \cdot 60\ 000$$

Die auftretenden Unterschiede zwischen tatsächlichem und errechnetem Wert ergeben sich durch die Wahl von n = 20, durch die sich zwei aufeinanderfolgende Texte um 3 000 Token unterscheiden. Wird n vergrößert, werden genauere Ergebnisse möglich; ideal wäre die Betrachtung bei n = 60 000, d.h. nach jedem neu hinzukommenden Token wird untersucht; eine praktische Auswertung würde dadurch allerdings erheblich kompliziert.

b. Sprunghöhe

Als Sprunghöhe wurde im vorhergehenden der Anteil der seltenen Wortformen des Textes definiert, wobei

$$F_{selt.\ Wf.} \leqq \frac{100}{B}$$

B-Wert	$F_{s.Wf.}$	Anzahl der Token der selt. Wf. = erwartete Sprunghöhe		gemessene Sprunghöhe
		abs.	% Text	
10	= 10	22970	38,28	41,57
15	= 6,7 [11]	20474	34,12	36,96
20	= 5	18029	30,05	33,52
25	= 4	16449	27,41	26,84
30	= 3,33	14349	23,90	30,32

Beachtung verdient das Ansteigen der tatsächlichen Sprunghöhe bei B = 30. Der Anteil der seltenen Wortformen muß, da $F_{s.Wf} \leqq \frac{100}{B}$, notwendigerweise sinken; der höhere Prozentsatz entsteht durch diejenigen Types, die, infolge der großen Grenze, hier bereits während und nicht wie sonst vor dem Sprung unter B liegen. Die Kurve M = f(B) weist zwischen B = 25 und B = 30 anscheinend einen Wendepunkt auf. Ein Vergleich der errechneten Sprunghöhen mit den Ergebnissen der anderen Teiltexte zeigt, daß bei B = 26 die größte Übereinstimmung herrscht. Bei diesem Wert enthält der Sprung mit größter Wahrscheinlichkeit nur die seltenen Wortformen und läßt damit die genauesten Aussagen zu.

Die gemessenen Ergebnisse in die Gleichungen des ersten Teils eingesetzt ergeben:

$$S = \frac{720 + 26 \cdot \sqrt{(26-B)^2} - \frac{T_n}{30\,000}}{T_n^{0,3}}$$

$$T_{Schnitt} \approx 10^4$$

$$V = \frac{T_k^{0,7} - T_{Schnitt}^{0,7}}{T_n^{0,7} - T_{Schnitt}^{0,7}}$$

$$T_k = T_n \cdot \sqrt[0,7]{V} \pm 5\,000$$

Am größten zur Verfügung stehenden Wortkorpus von Meier getestet, ergibt sich:

$$T_n \approx 11\,000\,000$$

Da Meier die Wortformen mit absoluter Häufigkeit zwischen 1 und 3 in einer Klasse zusammenfaßt und Prozentwerte dafür angibt, muß $F_{s.Wf.} \leqq 3$ gewählt werden, was einen B-Wert von 33,33 ergibt.

$$S = \frac{720 + 7{,}33 \cdot 26 - 366.667}{11\,000\,000^{0,3}} = 4{,}199$$

Meier gibt den Anteil der Wortformen zwischen 1 und 3 mit 3,71 an; die Differenz beträgt also 0,48 %.

Abb. 6 zeigt verschiedene andere Ergebnisse - einschließlich der Untersuchungstexte im Vergleich zur oben angeführten Formel zur Berechnung des Anteils der seltenen Wortformen.

Dabei ist:

$$\text{Gerade: } S = \frac{746 - \frac{T_n}{30\,000}}{T_n^{0,3}}$$

A - D: Untersuchungstexte zu 15 000, 30 000, 45 000 und 60 000 Token

1) : "The Captain's Daughter" (Puschkin) nach Herdan
2) : "Welt"-Text aus 12 500 Token
3) : "Neues Deutschland"-Text aus 41 500 Token (nicht identisch mit untersuchtem Text)
4) : auf 66 000 Token erweiterter Untersuchungstext

Die zahlenmäßige Verwandtschaft zwischen -a+1 und dem Potenzfaktor c in der Gleichung $V = T^c$ erscheint weiterer Untersuchungen wert. c hat im unteren Bereich (bei Texten mit ungefähr 10 000 Token) den Wert 0,9; er sinkt bei wachsender Textlänge und beträgt bei Meier $\approx$ 0,77. Nach der auf der Seite vorher angegebenen Formel wäre der Anteil der seltenen Wortformen gleich Null - bei B = 25 -, wenn

$$T_n \approx 22 \cdot 10^6$$

Es wäre denkbar, daß c und -a+1 bei dieser Textmenge den gleichen Wert annehmen.

Ziel dieser Arbeit war es, einen Vergleichsparameter zwischen Gesamt- und Teiltext zu ermitteln. Wir glauben, ihn im "V-Quotienten" gefunden zu haben, der Angaben über den Genauigkeitsgrad eines Auswahltexts -

selbstverständlich nur in quantitativer Hinsicht - zuläßt. Die Gleichungen des ersten Teils haben sich für das untersuchte Material als anwendbar erwiesen; die darin auftretenden Konstanten gelten mit größter Wahrscheinlichkeit für die Größenordnung des ausgewerteten Materials.
Ob und inwieweit sie für andere Texte zutreffen, muß weiteren Untersuchungen vorbehalten bleiben.

Für die konkrete Textaufnahme ergeben sich aus der Untersuchung folgende Punkte:

1. Eine Vervielfachung des Materials bringt nicht im gleichen Maße eine Vervielfachung der Genauigkeit.
2. Der gewählte Genauigkeitsgrad hängt vom Untersuchungsziel ab. Je seltener eine sprachliche Erscheinung im Gesamttext auftritt, desto größer muß der Wert der Wahrscheinlichkeit und damit der "V-Quotient" gewählt werden. Notwendigerweise wird dies bei grammatischen, syntaktischen oder stilistischen Fragestellungen der Fall sein, während Untersuchungen über Wortklassen, Wortfrequenzen usw. weniger hohe Wahrscheinlichkeit verlangen werden. Auch die Untersuchung seltener Sachgebiete stellt höhere Anforderungen an die Genauigkeit.
3. Bei gleichem Genauigkeitsgrad ist bei kleinen und großen Textmengen die relative Auswahlquote gleich. Von 2 Millionen Wörtern muß bei einem V-Quotienten bestimmter Größe der gleiche Prozentsatz aufgenommen werden wie bei 200 000 Wörtern.
4. Aussagen an Auswahltexten um oder unter 10 000 Wörtern können nicht als verbindlich für die Gesamttexte angesehen werden. Korpora dieser Größe müssen komplett aufgenommen werden.
5. Statt der verschwommenen Formulierungen "kann als ausreichend angesehen werden", "scheint genügend groß zu sein", "hat sich als hinreichend umfangreich erwiesen", die bei Untersuchungen an ausgewählten Texten auftauchen, sollte der Genauigkeitsgrad, die Signifikanz des Auswahltextes angegeben werden. Das hier vorgelegte Verfahren ist ein Vorschlag zur Ermittlung dieser Signifikanz.

Erläuterungen zu den Abbildungen

Abb. 1: Das Verhältnis zwischen erwartetem und tatsächlichem Auftreten der Token und der Abweichung bei einem Type (hier "ist")

Gerade 1: erwartetes Auftreten (Maximum 420 Token)
Kurve 2: tatsächliches Vorkommen (Maximum 420 Token)
Kurve 3: Abweichung zwischen erwartetem und tatsächlichem Vorkommen, berechnet nach Formel (1a), Seite 133. (Maximum 100%)

Abb. 2: Die Kurve der seltenen Wortformen nach der Näherungsgleichung

$$S = \frac{746 - \frac{T}{30000}}{T^{0,3}} \qquad \text{(Kurve 1)},$$

die Kurve der nicht seltenen Wortformen nach

$$NS = 95 - S \qquad \text{(Kurve 2)} -$$

beide nach Seiten 140 u. 141 , Formeln (2) und (3) und die von einem Text T_k gebildete Fläche (schraffiert).

Abb. 3 Neue Wortformen pro Texteinheit (= Seite). (Maximum 2000 Types)

Abb. 4: a - d Kennlinien verschieden langer Texte bei verschiedenen B-Werten.
Die x-Achse gibt die Textlänge an, die y-Achse den Prozentsatz der Worte, die unter der gesetzten Abweichung liegen. Von unter nach oben gelten für die vier Kurven eines Textes die B-Werte 10(%), 15(%), 20(%) und 25(%).

4a: Kennlinien des 15 000-Token-Textes
4b: Kennlinien des 30 000-Token-Textes
4c: Kennlinien des 45 000-Token-Textes
4d: Kennlinien des 60 000-Token-Textes

Auf die Darstellung des Nullpunktes wurde bei den Zeichnungen verzichtet.

Abb. 5: Die Verschiebung des Sprungpunktes bei wachsender Textlänge. Da die x-Achse zwischen 5 und 20 Einheiten variiert, ist die horizontale Verschiebung des Absprungpunktes scheinbar. Die Kennzeichnung der einzelnen Kurven erfolgt am Sprungpunkt; die Abweichung beträgt in allen vier Fällen B = 25. Kennung 1 - 4 wie bei 4 a - d.

Abb. 6: Anteil der seltenen Wortformen am Text.
Vergleich zwischen der errechneten Näherungsgleichung und verschiedenen Ergebnissen. (Siehe Seite 155)

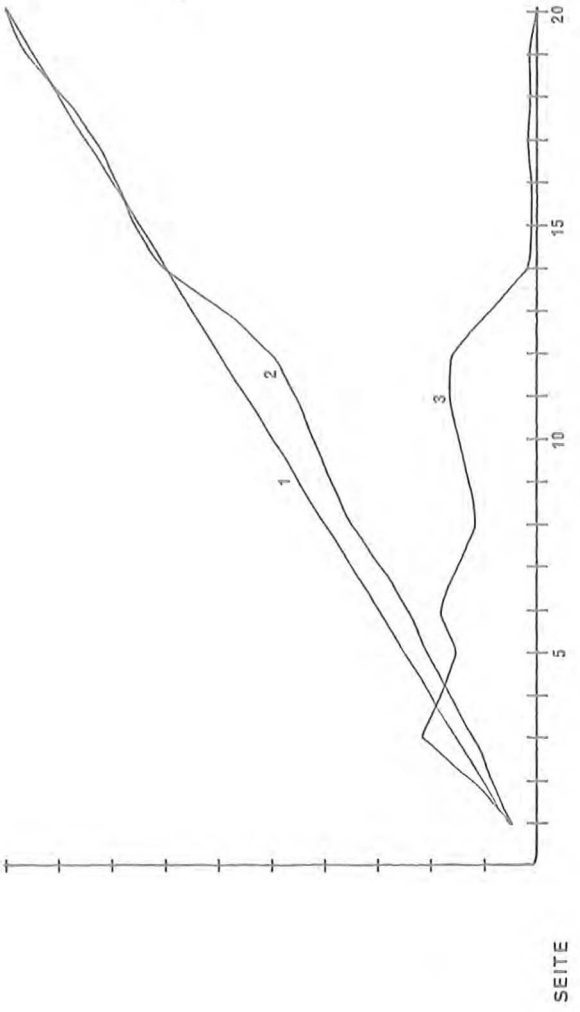


ABB. 1

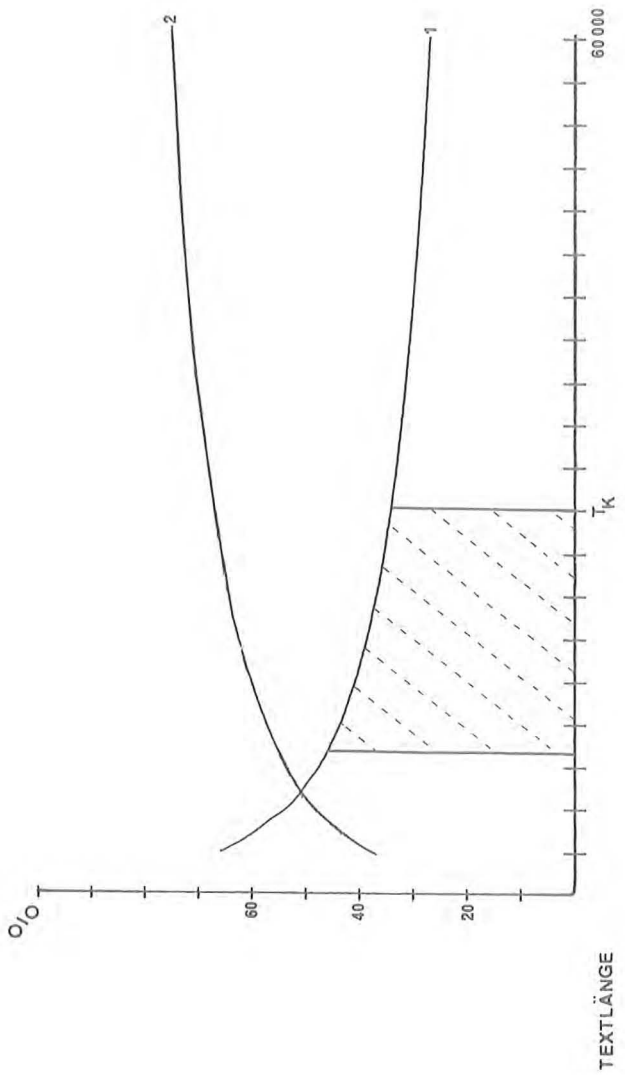
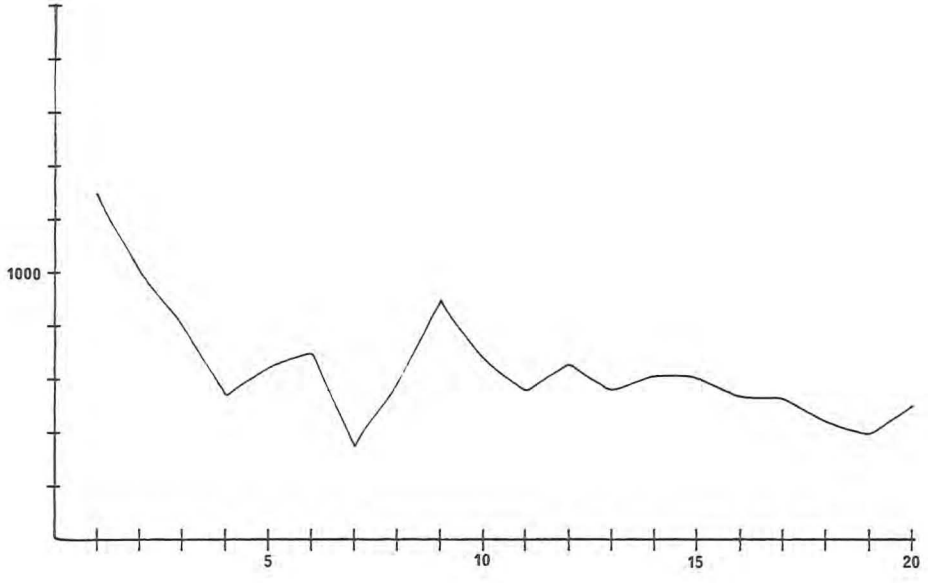


ABB. 2



SEITE

ABB. 3

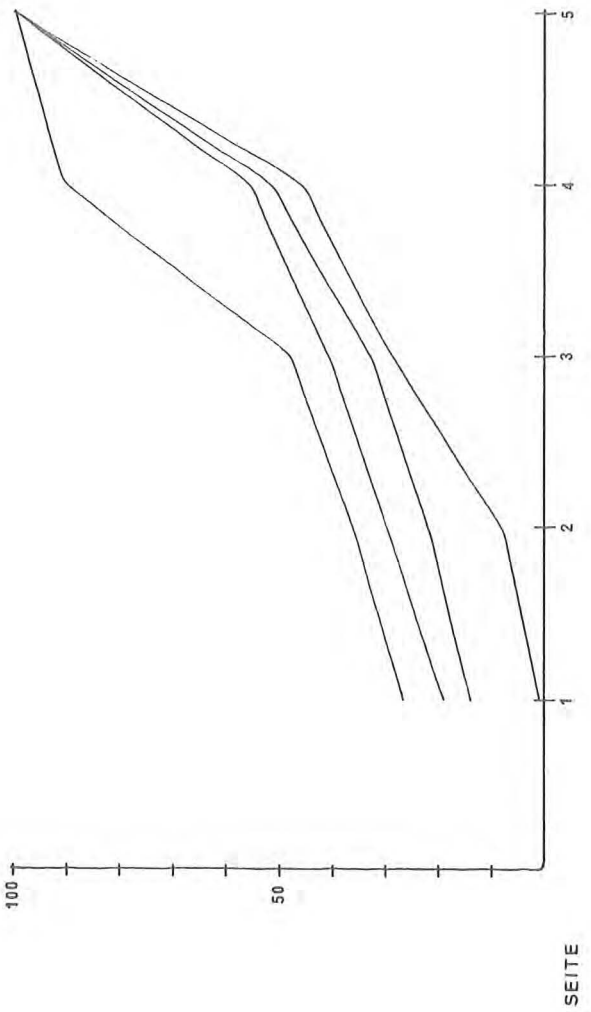


ABB. 4A

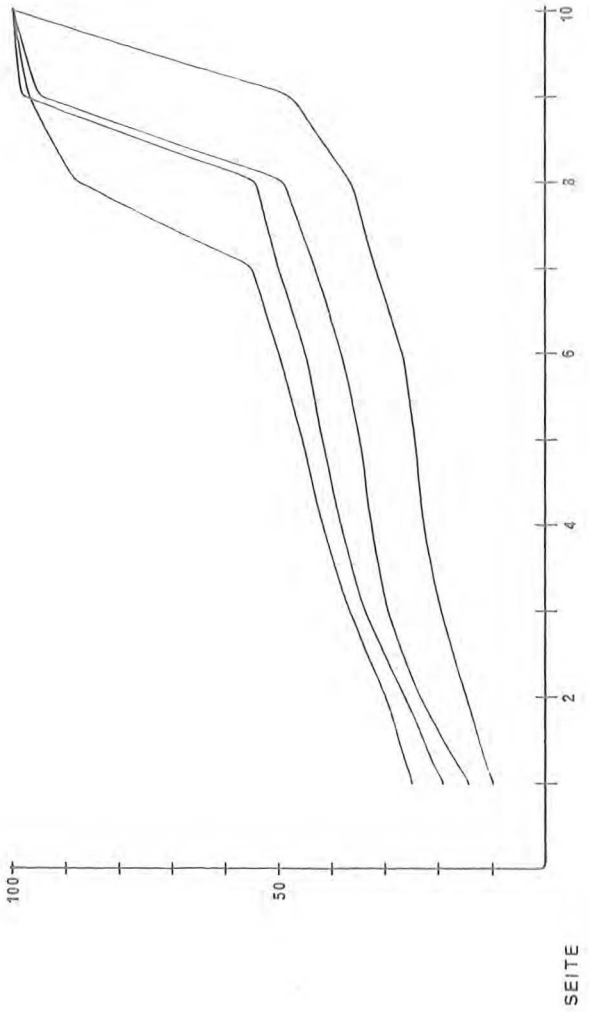


ABB.4B

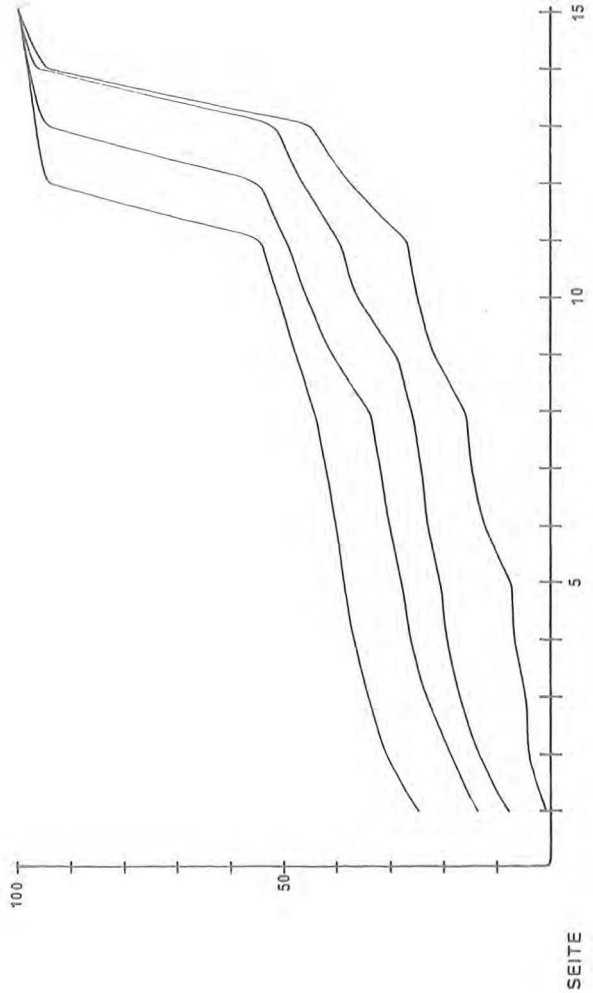


ABB. 4C

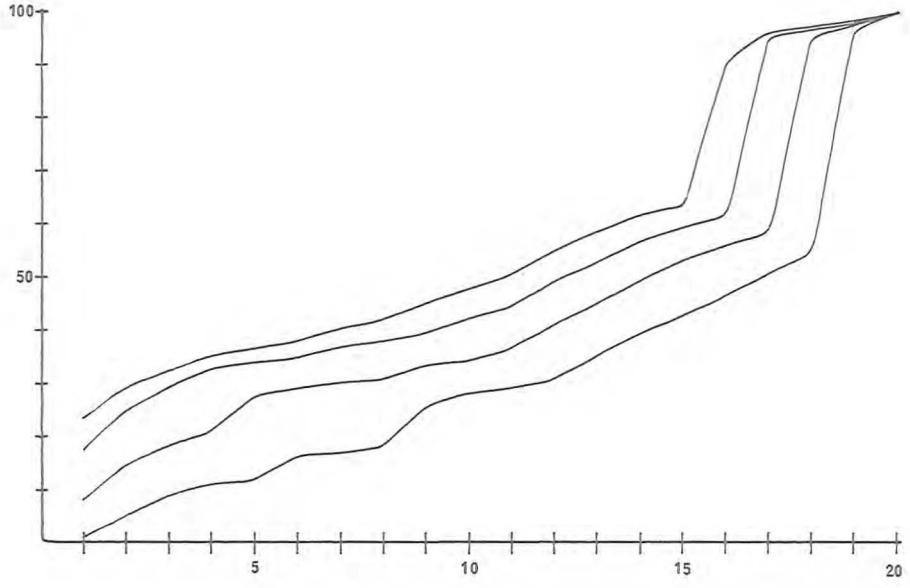


ABB. 4D

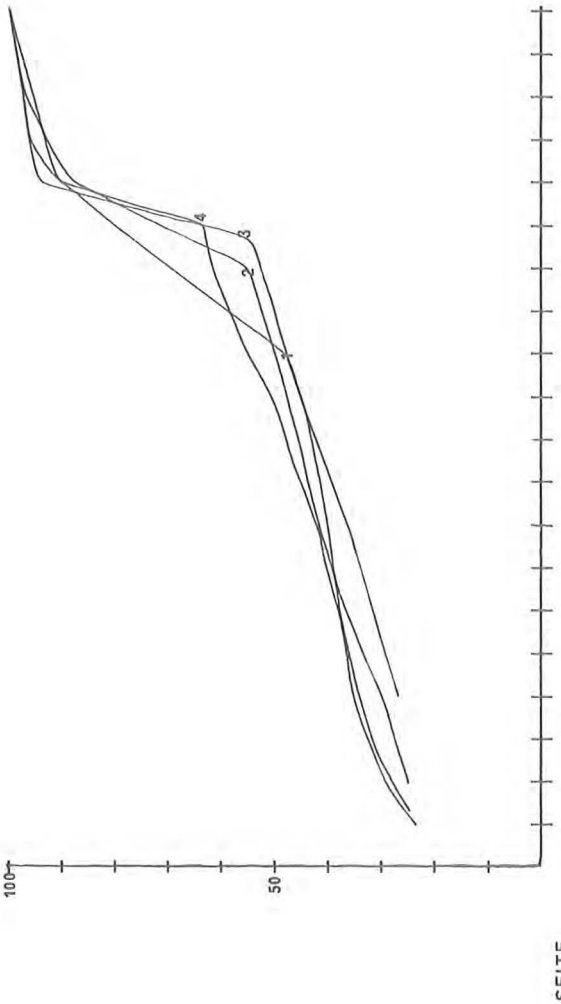


ABB. 5

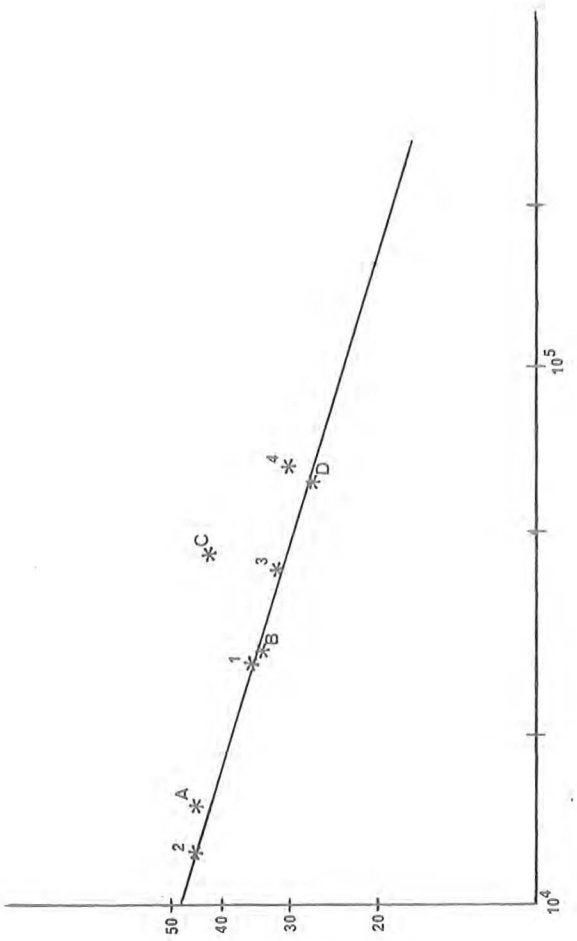


ABB. 6

Anhang I

Das Verhältnis von Vokabular und Textmenge

Token	Types	c-Wert bei $V = T^c$
3 000	1295	0.89475
6 000	2296	0.88895
9 000	3099	0.88285
12 000	3633	0.87235
15 000	4280	0.86895
18 000	4982	0.86855
21 000	5329	0.86195
24 000	5900	0.86040
27 000	6799	0.86455
30 000	7472	0.86475
33 000	8022	0.86355
36 000	8683	0.86395
39 000	9230	0.86335
42 000	9850	0.86335
45 000	10460	0.86315
48 000	10981	0.86295
51 000	11532	0.86255
54 000	11945	0.86115
57 000	12331	0.85985
60 000	12832	0.85925

Anhang II

Reihenfolge der häufigsten Wortformen im Vergleich zu anderen Untersuchungen

	untersuchtes Material	Eggers	Meier
1	der	die	die
2	die	der	der
3	und	und	und
4	in	in	in
5	den	das	zu
6	des	den	den
7	zu	ist	das
8	das	zu	nicht
9	von	von	von
10	für	des	sie
11	mit	sich	ist
12	auf	nicht	des
13	im	sie	sich
14	ist	es	mit
15	dem	dem	dem
16	es	eine	daß
17	daß	als	er
18	sich	ein	es
19	eine	mit	ein
20	nicht	im	ich
21	auch	auf	auf
22	werden	daß	so
23	sie	er	eine
24	an	auch	auch
25	aus	für	als

Anhang III

Belegung der Häufigkeitsklassen

absolute Hfgk.	Anzahl d. Types	Anzahl d. Token	Token kumuliert
1	7925	7925	7925
2	1919	3838	11763
3	862	2586	14349
4	525	2100	16449
5	316	1580	18029
6	222	1332	19361
7	159	1113	20474
8	119	952	21426
9	86	774	22200
10	77	770	22970
11	56	616	23586
12	55	660	24246
13	41	533	24779
14	28	392	25171
15	33	495	25666
16	32	512	26178
17	22	374	26552
18	25	450	27002
19	23	437	27439
20	17	340	27779
21	19	399	28178
22	12	264	28442
23	10	230	28672
24	7	168	28840
25	11	275	29115

Anmerkungen zu Teil III

1) Sämtliche Aussagen betreffen nur quantitative Eigenschaften des Textes.

2) Herdan, G.: The advanced Theory of Language as Choice and Chance. Berlin 1966

3) Pfanzagl, J.: Allgemeine Methodenlehre der Statistik I. Berlin 1966, S. 144

4) Herdan a.a.O., S. 96

5) Da die Abweichung bei wachsendem k ($\rightarrow$ n) nach Null geht, bleibt die Richtung von A unberücksichtigt; die Umkehrung der Formel ist möglich.

6) Bemerkungen zu dieser und den folgenden Abbildungen siehe "Erläuterungen zu den Abbildungen", Seite 157.

7) Hierfür empfiehlt es sich, auch die Richtung der Abweichung zu berücksichtigen, um den Sprung zu verdeutlichen. Etwa:
$H = \left[A(w,T_k) - A(w,T_n)\right] \cdot k$

8) Selbstverständlich treten auf dieser Seite noch andere textsignifikante Wörter auf.

9) Meier, Helmut: Deutsche Sprachstatistik, Hildesheim 1964.

10) Eggers, Hans: Beobachtungen zur Häufigkeit deutscher Wortformen. In: Wirkendes Wort 1967, Heft 2.

11) Natürlich ist es unmöglich, Häufigkeiten zwischen ganzen Zahlen zu ermitteln; es muß die nächste ganze Zahl als Grundlage genommen werden. Außerdem muß berücksichtigt werden, daß zwischen den beiden Textteilen, die den Sprung begrenzen (k und $k + \frac{N}{n}$) auch noch andere Wortformen als die seltenen unter B sinken können.

Literatur

Eggers, Hans : Beobachtungen zur Häufigkeit deutscher Wortformen. In: Wirkendes Wort 1967, Heft 2

Herdan, G. : The advanced Theory of Language as Choice and Chance. Berlin 1966

Kaeding, Fr.W. : Häufigkeitswörterbuch der deutschen Sprache. 1897-98

Meier, Helmut : Deutsche Sprachstatistik. Hildesheim 1964

Pfanzagl, J. : Allgemeine Methodenlehre der Statistik. I. Berlin 1966