Erschienen in: Moser, Hugo/Glinz, Hans/Grebe, Paul/von Polenz, Peter (Hrsg.): Foschungsberichte des Instituts für deutsche Sprache Band 2. - Mannheim: Institut für Deutsche Sprache, 1968. S. 1-8. (Forschungsberichte des Instituts für deutsche Sprache 2)



# 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 nicht 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 Schulte-Tigges 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 Eggers und seinen Mitarbeitern von der Universität Saarbrücken sowie dem Institut für Phonetik und Kommunikationsforschung in Bonn, besonders dessen Direktor, Professor Dr. Gerold Ungeheuer, und Dr. Dieter Krallmann. 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 repräsentatives 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 gesprochener Texte, das auf einen endgültigen Umfang von mindestens 600.000 Wörtern angelegt ist, wird zur Zeit unter Leitung von Professor Dr. Hugo Steger (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.