345 Fr.

FORSCHUNGSBERICHTE

DES

INSTITUTS FÜR DEUTSCHE SPRACHE

herausgegeben

gemeinsam mit Hans Glinz, Paul Grebe und Peter von Polenz

von Hugo Moser

Schriftleitung: Ulrich Engel

3

April 1969

Das Mannheimer Corpus

von Ulrich Engel

Da sich die in diesem Forschungsbericht enthaltenen Beiträge ausdrücklich oder implizit mit den vom Institut für deutsche Sprache zusammengestellten und für weitere elektronische Bearbeitung auf Magnetband gespeicherten Texten befassen, da scheint es angebracht, näher auf Geschichte und Motivation des "Mannheimer Corpus" einzugehen.

Dieses Corpus (s. Seite 76 - 78) beruht auf Vorschlägen der wissenschaftlichen Mitarbeiter des Instituts, die in mehreren Sitzungen vom Wissenschaftlichen Rat und den dafür zuständigen Kommissionen (Dokumentation, Grunddeutsch) erörtert und ergänzt wurden.

Es wurde eine Auswahl angestrebt, die als repräsentativ angesehen werden konnte für die deutsche Gegenwartssprache (seit 1945) in ihrer geschriebenen Form, wobei von regionalen und sozialen Dialekten und anderen Sonderformen mit beschränktem Geltungsbereich abgesehen wurde.

Damit war von vornherein klar, daß das Corpus nicht auf die sogenannte schöne Literatur beschränkt werden durfte. Im ganzen schälten sich drei große Bereiche heraus: Neben der schönen Literatur einschließlich ihrer Trivialformen vor allem die Fachliteratur, die sowohl in Einzelbänden als auch in mehreren Zeitschriften erfaßt wurde; schließlich die Sprache der politischen Nachrichten, wie sie sich im wesentlichen auf Seite 1 der Tageszeitungen finden. Die "Bild Zeitung" freilich ist als ein Phänomen sui generis zu betrachten, zweifellos enthält sie nicht einfach einen spezifischen Aspekt von Zeitungssprache.

Eine wichtige Rolle spielte die Frage, was tatsächlich gelesen wird. Aus dieser Fragestellung unter anderem[1)] erklärt sich die Einbeziehung der Trivialliteratur mit 3 Werken, die starke Bevorzugung populärwissenschaftlicher Werke gegenüber der Fachprosa im engeren Sinne, schließlich die Aufnahme eines halben Jahrganges der Bild Zeitung.

Deutsche Textbibliothek

Maschinell gespeicherte Texte des Instituts für deutsche Sprache

Bergengruen, Werner, Das Tempelchen, Erzählung, Arche, Zürich, Nymphenburger Verlagshandlung, München (C. 1950 Peter Schifferli, Verlags AG Die Arche, Zürich).

Böll, Heinrich, Ansichten eines Clowns, Roman, Kiepenheuer und Witsch, Köln - Berlin, 113.- 157. Tausend August 1964 (1. - 28. Tausend Mai 1963).

Frisch, Max, Homo Faber, Ein Bericht, Bibliothek Suhrkamp, Bd. 87, 161.-180. Tausend 1966 (C. 1957 Suhrkamp, Frankfurt/Main).

Grass, Günter, Die Blechtrommel, Roman, Firscher, Frankfurt/Main und Hamburg, 323.- 372. Tausend Mai 1964 (1.-50. Tausend September 1962) (C. 5. und 6. Auflage August 1960 Luchterhand, Darmstadt - Berlin - Neuwied).

Johnson, Uwe, Das dritte Buch über Achim, Roman, Suhrkamp, Frankfurt/Main 1961, 16.-20. Tausend (C.1961).

Mann, Thomas, Die Betrogene, Erzählung, S. Fischer, Frankfurt/Main, 16.- 20. Tausend 1954 (C. 1953).

Strittmatter, Erwin, Ole Bienkopp, Roman, Sigbert Mohn, Gütersloh (C. 1963 Aufbau-Verlag, Berlin W8).

Jung, Else, Die Magd vom Zellerhof, Kelter Heimat-Roman Bd. 41, Martin Kelter, Hamburg-Wandsbeck (o.J.).

Pinkwart, Heinz, Mord ist schlecht für hohen Blutdruck, Kriminalroman, Goldmann, München (C. 1963).

Stauffen, Pia, Solange dein Herz schlägt, Juwelen-Roman Nr. 748, Pabel, Rastatt (Baden) (o.J.).

Bamm, Peter, Ex Ovo, Essays über die Medizin, Deutsche Verlags-Anstalt, Stuttgart, 63.- 65. Tausend 1963 (C. 1956).

Bollnow, Otto Friedrich, Maß und Vermessenheit des Menschen, Philosophische Aufsätze, Neue Folge, Vandenhoeck & Ruprecht, Göttingen und Zürich (C. 1962).

Gail, Otto Willi und Petri, W., Weltraumfahrt, Physik - Technik - Biologie, 2., völlig neubearbeitete Auflage des Werkes Physik der Weltraumfahrt , 1947, Hanns Reich, München (C. 1958).

Grzimek, Bernhard, Serengeti darf nicht sterben, Ullstein, Berlin, 131.- 141. Tausend April 1963 (1.-30. Tausend September 1959).

Heimpel, Hermann, Kapitulation vor der Geschichte?, 3., vermehrte Auflage, Vandenhoeck & Ruprecht, Göttingen, 13.-18. Tausend 1969 (C. 1956).

Heisenberg, Werner, Das Naturbild der heutigen Physik, rde 8, Rowohlt, Hamburg, 124. - 128. Tausend September 1966 (1.- 40. Tausend Dezember 1955), bis S. 46.

Jaspers, Karl, Die Atombombe und die Zukunft des Menschen, Piper, München, 37.- 44. Tausend 1962 (C. 1958).

Jungk, Robert, Die Zukunft hat schon begonnen, Amerikas Allmacht und Ohnmacht, neue erweiterte Ausgabe, Scherz, Bern - München - Wien (C. 1952).

Pörtner, Rudolf, Die Erben Roms, Städte und Stätten des deutschen Früh-Mittelalters, Econ, Düsseldorf - Wien, 41.- 70. Tausend 1965 (1.- 40. Tausend 1964).

Staiger, Emil, Grundbegriffe der Poetik, Atlantis, Zürich - Freiburg/Breisgau 1966, 7. Auflage (C. 1946).

Ullrich, Fritz, Wehr Dich Bürger! Aktuelle Rechtsschutzfibel, Gieseking, Bielefeld (C. 1960).

Heuß, Theodor, Erinnerungen, 1905-1933, Wunderlich, Tübingen, 5. Auflage, 71.- 85. Tausend Mai 1964 (1. Auflage September 1963).

"Frankfurter Allgemeine Zeitung", D-Ausgabe, 19.1.1966 - 17.2.1966, jeweils die erste Seite (ohne Leitartikel).

"Die Welt", Ausgabe D***, 1.12.1965 - 18.2.1966, jeweils die erste Seite.

"Bild der Wissenschaft", hg. von Prof. Dr. Heinz Haber in der Deutschen Verlags-Anstalt, Stuttgart, Heft 1, 2 und 3/1967 (jedes Heft enthält 5 Artikel).

"Studium Generale", Schriftleitung G.G. Grau, Springer, Berlin - Heidelberg - New York, Heft 12/1966 (das Heft enthält 6 Artikel).

"Urania", hg. vom Präsidium der URANIA (Gesellschaft zur Verbreitung wissenschaftlicher Kenntnisse) und dem Deutschen Kulturbund, Urania, Leipzig - Jena - Berlin, Heft 11/1966 und 1/1967 (jedes Heft enthält 14 Artikel).

"Bild Zeitung" 7 Monate (Januar-Juli 1967), im Turnus 1.Tag/1.Seite, 2. Tag/2.Seite ... 7. Tag/1.Seite usw.

Auch der regionale Gesichtspunkt wurde berücksichtigt. Die Autoren vor allem der Bereiche "Schöne Literatur" und "Trivialliteratur" stammen aus verschiedenen Teilen des deutschen Sprachraums, mit Max Frisch ist auch die deutschsprachige Schweiz vertreten, die Aufnahme eines österreichischen Autors ist vorgesehen. Da die Möglichkeit, daß sozial und politisch verschiedene Systeme auch zu sprachlicher Differenzierung führen, nicht von der Hand zu weisen ist[2)], wurde mit dem Ole Bienkopp auch das Werk eines anerkannten Schriftstellers aus der DDR aufgenommen.

Schließlich war auf die Erfassung möglichst verschiedener Stilarten zu achten. Nachdem geschriebene Gegenwartssprache definiert worden war als Sprache der Gesamtheit der nach 1945 entstandenen Werke, mußte den hier offenkundig vorhandenen Stilunterschieden Rechnung getragen werden. Es wäre unzulässige Simplifizierung, stilistische Spezifika einfach vom Alter des Autors abhängig zu machen oder gar schlicht auf den Unterschied der Generationen zurückzuführen. Es kann aber nicht übersehen werden, daß eine Reihe vorwiegend älterer Autoren einen weitgehend an der deutschen Klassik orientierten Stil pflegt, während die Jüngeren bewußt mit der Schultradition brechen - faktisch oft viel weniger, als offenbar beabsichtigt ist - und neue, freiere, elastischere, der Alltagssprache näherstehende Stilformen sich heranzubilden beginnen. Auf dem Hintergrund solcher Beobachtungen mag es verständlich werden, daß (neben Frisch, Grass, Johnson) Thomas Mann und Werner Bergengruen mit Spätwerken aufgenommen werden.

Natürlich waren auch Beschränkungen notwendig. Von der maschinellen Erfassung wie von der linguistischen Auswertung her waren äußere Grenzen gesetzt: der Grad der Zuverlässigkeit der Ergebnisse muß in einem vernünftigen Verhältnis zu der aufgewendeten Arbeit stehen. Ein Corpus im Umfang von insgesamt etwa 1,6 Millionen Wörtern - rund das Dreißigfache von Max Frischs "Homo Faber" - konnte einerseits als ausreichend gelten, andererseits erlaubt es noch eine einigermaßen vollständige Auswertung.

Es war schon die Rede davon, daß Schwerpunktbildung erforderlich war. Dies gilt vor allem für den Bereich der wissenschaftlichen Literatur. Hier wurde, aus schon dargelegten Gründen, eng Fachbezogenes ausgeschieden, weil hier - zumeist

beim Wortschatz - Sonderbildungen zu erwarten waren, die für das Gesamtbild der deutschen Gegenwartssprache von zweitrangiger Bedeutung sind. Außerdem wurde aber auch, und zwar mit ausdrücklicher Zustimmung der Vertreter des Goethe-Instituts, für dessen Arbeit dieser Zweig besonders bedeutsam ist, die Gesamtmenge der Lehr- und Fachbücher weggelassen. Selbst durch die Unterrichtsliteratur sehr niederen Niveaus käme nämlich schon ein umfangreicher Wortschatz sehr spezieller Art in das Corpus, während die generellen lexikalischen und syntaktischen Merkmale der Fach- und Wissenschaftssprache ebensogut in den (vorwiegend) populärwissenschaftlichen Werken und den berücksichtigten Zeitschriften vertreten sind.

Daß von den beiden Zeitungen "Die Welt" und "Frankfurter Allgemeine Zeitung" - die "Bild Zeitung" ist in mehrfacher Hinsicht als Sonderfall zu betrachten - nur die erste Seite aufgenommen wurde, hat wohldurchdachte Gründe. Es ging ja von Anfang an nicht um "Zeitungssprache", ein ohnehin höchst heterogenes Gebilde, dessen Legitimation als Gegenstand linguistischer Forschung noch erbracht werden müßte, sondern es ging um die Sprache der politischen Nachrichten, dessen Untersuchung Peter von Polenz dezidiert gefordert hatte[3]. Diese Sprache findet sich zwar auch an anderen Stellen normaler Tageszeitungen, aber es ist legitim, die erste Seite der Zeitung als stellvertretend für den sprachlichen Gesamtbereich herauszugreifen.

Es stellt sich auch die Frage, in welchem Umfang die einzelnen Werke zu erfassen seien. Statistiker empfehlen in solchen Fällen immer möglichst viele kleinere Teiltexte; ". . . daß statistisch gesehen 10 Prozent aus 100 Titeln immer besser sind als 100 Prozent aus 10 Titeln", betont auch Manfred Hellmann in Abschnitt 2.5. seines Berichts. Diese Forderungen bestehen grundsätzlich zu Recht, wenn es in erster Linie darum geht, in dem Auswahlcorpus eine größere Grundgesamtheit möglichst "maßstabsgetreu" zu repräsentieren. Wenn wir uns doch zur Aufnahme der ungekürzten Texte entschlossen haben, so hauptsächlich auf Grund der Erwägung, daß unser Corpus für möglichst viele beliebige Fragestellungen geeignet sein sollte. Dazu gehören aber auch mögliche kompositorisch bedingte Erscheinungen, die nur im weitesten Kontextzusammenhang untersucht werden können und deshalb in Auswahltexten verlorengehen könnten.

Die gesamte Diskussion läuft letztlich auf die Frage hinaus, was wir unter der Repräsentativität eines Corpus verstehen. Daraus wiederum ergeben sich zwei Teilfragen:

1) Wofür soll das Corpus repräsentativ sein?
2) Im Hinblick worauf soll es repräsentativ sein?

1) Wir betrachten als Grundgesamtheit, die vom Corpus abgebildet werden soll, die gemeindeutsche (interregionale, intersoziale, überfachliche) Gegenwartsprosa. Bei Anlegung streng statistischer Maßstäbe müßten möglichst viele Merkmale dieser Grundgesamtheit im Corpus vertreten sein, und zwar jeweils entsprechend der Häufigkeit ihres Vorkommens.

Es ist aber bisher noch kein Verfahren entwickelt worden, das die Gewinnung eines solcherart repräsentativen Corpus ermöglicht, ohne daß die Grundgesamtheit en détail bekannt wäre (wahrscheinlich wird ein solches Verfahren auch nie bis zur Funktionsreife gedeihen). Wie die Fülle des deutschen Prosaschrifttums unserer Zeit systematisch erfaßt werden könnte (man denke nur an die Mengen täglich erscheinender Zeitungen, auf die Hellmann hinweist), ist ein vorderhand gänzlich ungelöstes Problem. Das hängt weitgehend zusammen mit der Tatsache, daß eine praktikable Typik des deutschen Schrifttums nicht existiert. Der Kernsche Entwurf bietet ein wertvolles, einleuchtendes Instrumentarium für Textbeschreibungen und damit auch für eine Typik der deutschen Gegenwartssprache. Kriterien für die Typenbildung bietet er nicht, er will das auch gar nicht: dies zeigt der dritte Teil seiner Studie, in der präexistente "Typen" durch Kombination mehr oder weniger zahlreicher Merkmale charakterisiert werden. Weitgehend ist man sich über bestehende Typen einig (die literarischen Gattungen und anderes werden meist unreflektiert mit verwendet), ohne ihre Konstituierung begründen zu können: eine allgemeine, verbindliche Typik muß noch geschaffen werden.

Hätten wir diese Typik, hätten wir mithin ein allumfassendes "Schubkastensystem", in das jeder Text auf Grund exakt feststellbarer Merkmale eingeordnet werden könnte (wobei simultane Einordnung in verschiedene "Schubkästen" inbegriffen wäre), so könnten wir tatsächlich ein quantitativ getreues Abbild der Grundgesamtheit liefern: ein übersichtliches Corpus, in dem alle "Typen" des Gegenwartsschrifttums

entsprechend der Häufigkeit ihres Vorkommens vertreten wären.

Es fragt sich, was damit gewonnen wäre. Die wöchentlichen Bestsellerlisten des "Spiegel" verzichten mit gutem Grund auf die Einbeziehung der absoluten Bestseller, der Lehrbücher jeglicher Art und Provenienz. Das rein quantitative verkleinerte Abbild einer strikt typisierten Gegenwartsliteratur enthielte einen alles andere überwiegenden Teil an didaktischer Literatur, der in seiner Monotonie nur wenige Aufschlüsse verspräche. Schon deshalb scheint uns ein Ausgehen von den Auflagenziffern nicht diskutabel. Das deutsche Gegenwartsschrifttum kann durch bloße Aufreihung alles in deutscher Sprache Gedruckten nicht adäquat erfaßt werden. Neben dem Ausstoß der Druckereien muß mindestens auch der Grad der Wertschätzung beim lesenden Publikum, eingeschlossen die mehr oder weniger etablierten, von Kritikern oder von der Schule verbreiteten Wertnormen, berücksichtigt werden. Zahlreiche weitere Gesichtspunkte kommen hinzu. Es scheint mir weit eher vertretbar, von einer solchen auf Grund einer Vielzahl relevanter Kriterien "gewichteten", wenngleich quantitativ nicht so exakt abgegrenzten Gesamtmenge auszugehen, als von einer bloßen Summation des irgendwo und irgendwie Gedruckten. Für diese mehrfach gewichtete Gesamtmenge allerdings soll unser Corpus repräsentativ sein insofern, als es seine wesentlichen Merkmale ebenfalls enthält. Das ist, wie die bisherigen grammatischen Untersuchungen erwiesen haben (es wurden zum Vergleich und zur Ergänzung weitere Werke beigezogen) in hohem Maße der Fall. Freilich fehlen viele sprachlichen Sonderausprägungen. Man muß aber bedenken, daß dieses Corpus in erster Linie für die Untersuchung grammatisch-syntaktischer Erscheinungen erstellt wurde, die sich zum Teil durch sehr verschiedenartige Texte hindurch nur unwesentlich ändern. Ein speziell für Wortschatzuntersuchungen zusammengestelltes Corpus[4] müßte naturgemäß ganz anders aufgebaut werden. Dies führt uns zu der Frage:

2) Im Hinblick worauf soll das Corpus "repräsentativ" sein? Die Antwort auf diese Frage hängt eben nicht bloß von der definierten Gesamtmenge ab, deren Merkmale sich im Corpus wiederfinden sollen, sondern in noch höherem Maße von den zu untersuchenden Erscheinungen. Darüber liegen im Institut umfangreiche Erfahrungen vor. Während sich etwa das Corpus für Satzstrukturen, Wortstellung und bestimmte

Tempora als viel zu umfangreich erwies, so daß jeweils nur mit Teilmengen gearbeitet wurde, mußten für eine relativ seltene Erscheinung wie das Passiv in erheblichem Umfang weitere Texte beigezogen werden. Strenggenommen müßte für jede spezielle Fragestellung ein eigenes Corpus zusammengestellt werden. Wie ein solches Corpus jeweils beschaffen sein muß, ist oft nicht im voraus auszumachen, sondern ergibt sich vielfach erst als Teilergebnis der Untersuchung. Insofern können die Mannheimer Arbeiten zur deutschen Grammatik Hinweise für spätere weiterführende Untersuchungen ergeben.

Wo es dann um die Gewinnung von Teilmengen aus dem einmal festgelegten Corpus geht, soll freilich möglichst exakt verfahren werden. Das von Werner Müller dargelegte Verfahren ist für solche Zwecke nützlich. Billmeiers Vorschlag (im Forschungsbericht Nr. 2) ist zunächst für Wortschatzuntersuchungen konzipiert; die Anwendung auf grammatische Fragestellungen dürfte erheblich komplizierter sein.

Wir fassen zusammen: Die Erstellung eines Corpus, das die gesamte deutsche Gegenwartsprosa exakt abbildet, ist heute noch auf lange Sicht unmöglich, und sie wäre auch aus verschiedenen Gründen kaum vertretbar. Da sich die jeweilige Fragestellung als wichtige Konstante bei der Corpuskonstitution erwiesen hat, kann es ein allgemein gültiges Corpus auch gar nicht geben, man müßte es denn in allen Teilen so umfangreich anlegen, daß die Auswertung alle Grenzen der Wirtschaftlichkeit sprengen würde. Das Corpus des Instituts für deutsche Sprache war immer nur als Rahmenkonzeption zu verstehen; es kann bei Bedarf ebenso erweitert wie nur partiell für die Auswertung herangezogen werden. Würde man heute noch einmal beginnen, so würde man gewiß manches ändern. Vor vier Jahren fehlten viele der heute vorliegenden Erfahrungen. Im ganzen freilich würde sich wahrscheinlich ein nicht allzu stark abweichendes Bild ergeben. Wir halten dieses Corpus für repräsentativ für die deutsche Gegenwartssprache in einem viel komplexeren als dem rein quantitativen Sinn; nämlich insofern, als die aus ihm gewonnenen grundlegenden Befunde mit gebührender Vorsicht so verallgemeinert werden können, daß sie auch zugleich Aussagen über die deutsche Gegenwartsprosa zulassen. Mehr war nie verlangt, mehr war auch nie behauptet worden.

Anmerkungen

1) Für die Auswahl der einzelnen Texte waren jeweils mehrere verschiedene Gesichtspunkte maßgebend.

2) Im allgemeinen überschätzt man freilich die sprachlichen Folgen politisch-sozialer Grenzziehungen. Vgl. dazu Hugo Mosers illustrative und kritische Schrift "Sprachliche Folgen der politischen Teilung Deutschlands", 1962 = Beiheft zum "Wirkenden Wort", 3; auch "Das Aueler Protokoll". Deutsche Sprache im Spannungsfeld zwischen West und Ost, Düsseldorf (Schwann) 1964.

3) Peter von Polenz, zur Quellenwahl für Dokumentation und Erforschung der deutschen Sprache der Gegenwart, in: Satz und Wort im heutigen Deutsch = Sprache der Gegenwart, Band I, S. 363-378.

4) Wie das von der Bonner Außenstelle erarbeitete Corpus, das der Ermittlung der (vorwiegend lexikalischen) sprachlichen Besonderheiten in beiden Teilen Deutschlands dient.