# Repräsentativität: von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung.

*Burghard Rieger*

**1.**

> Die Sprache liegt nur in der verbundenen Rede, Grammatik und Wörterbuch sind kaum ihrem todten Gerippe vergleichbar. Die bloße Vergleichung selbst dürftiger und nicht durchaus zweckmäßig gewählter Sprachproben lehrt daher viel besser den Totaleindruck des Charakters einer Sprache auffassen, als das gewöhnliche Studium der grammatischen Hülfsmittel. (...) Freilich fährt dies in eine mühevolle, oft ins Kleinliche gehende Elementaruntersuchung, es sind aber auch lauter in sich kleinliche Einzelheiten, auf welchen der Totaleindruck der Sprache beruht, und nichts ist mit dem Studium derselben so unverträglich, als bloß in ihnen das Große, Geistige, Vorherrschende aufsuchen zu wollen. (Humboldt, 1829, S. 186/200)

Es ist eines der schönen Resultate, mit dem die allemal vorteilhafte Lektüre der Werke Wilhelm v. Humboldts verbunden ist, daß sich – im „Labyrinth der Sprachkunde“ fortschreitend – für nahezu beliebige (überfällige oder auch nur vermeintliche) Neuansätze und (notwendige oder auch nur vorgebliche) Rückbesinnungen zumindest zitierbare Belegstellen finden lassen. Was bekanntermaßen für die spezifisch deutsche Entwicklung der strukturalen Sprachwissenschaft gilt, die als ‚inhaltsbezogene Grammatik' an Bedeutung gewann, und was – nicht weniger geläufig – für jenen Zweig des vornehmlich amerikanischen Strukturalismus zutrifft, der mit der Kompetenz-orientierten Entwicklung generativ-transformationeller Grammatiktheorien verbunden ist, kann auch in Anspruch genommen werden von dem inzwischen wieder verstärkt einsetzenden Bemühen um eine Performanz-orientierte Sprachwissenschaft, die – eher an der tatsächlichen Sprachverwendung als am idealisierten Sprachsystem interessiert – zunehmend empirisch-textanalytisch arbeitet und weniger formal-satzgenerierend ausgerichtet ist.

Die von Humboldt genannten Aspekte und deren zumindest angesprochener Zusammenhang können dabei als Bestimmungsstücke auch jener bis heute nicht abgeschlossenen Neuorientierung gelten, die – wenngleich mit unterschiedlichen Akzentuierungen und Terminologien – zu Beginn der 70er Jahre sich abzuzeichnen begann (Bellert 1970; Brinker 1971; van Dijk 1972, Dressler 1972; Hartmann 1971; Petöfi 1972, Rieger 1972; Wunderlich 1971; etc.).

Die programmatische wie praktische Anerkennung der „verbundenen Rede“ (*Text*) als der primären Erscheinungsform allen sprachlichen Handelns und Verhaltens betont den Vorrang der in tatsächlichen Kommunikationssituationen von wirklichen Sprechern/Hörern verwendeten Sprache (*Performanz*) gegenüber jenen abstraktiven „kaum ihrem Todtengerippe vergleichbaren“ formalen Strukturbeschreibungen und Systemen, wie sie die von aller individuellen, situativen, sozialen, historischen, etc. Variabilität weitgehend absehenden Regelwerke in „Grammatik“ und/oder „Wörterbuch“ (Lexikon) nur reduziert bieten.

Die im kommunikativen Handlungsschema funktionierende Produktion und Rezeption

von bedeutungskonstituierenden Texten (parole) wird dabei als durch jeweilige Aktualisierung eben jener abstrakten Systeme und Strukturzusammenhänge (langue) gedacht. Das soll umgekehrt erlauben, aus der Analyse und „Vergleichung“ ausschnitthaft „gewählter Sprachproben“ (*Korpora*) nicht nur die diese Teilmengen strukturierenden (mehr oder weniger ausgeprägten) Regularitäten zu bestimmen, sondern aus diesen auch auf die möglicherweise übergreifenden Strukturenzusammenhänge rückzuschließen, „auf welchen der Totaleindruck der Sprache beruht“. Dazu wird „eine mühevolle, oft ins Kleinliche gehende Elementaruntersuchung“ notwendig, die freilich – selbst wenn sie große Mengen „lauter in sich kleinlicher Einzelheiten“ berücksichtigt – heute so mühevoll nicht mehr zu sein braucht, sofern nämlich hierzu das Hilfsmittel der elektronischen Datenverarbeitung (*Computer*) gezielt eingesetzt werden kann.

**1.1.** Freilich lassen *Computer*-unterstützte Analysen *performativ* realisierter Sprache anhand „selbst dürftiger und nicht durchaus zweckmäßig gewählter“ *Korpora* nicht schon die systematische Kontrolle von Hypothesenbildungen zu, in denen etwa Schlüsse formuliert werden von den in den Textkorpora beobachteten Strukturen der Sprachverwendung auf deren vermutete Voraussetzungen oder Zusammenhänge im wie immer bestimmten Sprachsystem oder dessen Teilbereichen. Die allgemeine Bedingung für deren Beurteilung und Bewertung besteht vielmehr in der Möglichkeit, aufgrund der innerhalb der Thematik der linguistischen Basistheorie entwickelten Hypothesen weniger dürftige, dabei aber durchaus zweckmäßig gewählte Sprachausschnitte (Stichproben) aus der Menge aller möglichen sprachlich realisierten Äußerungen (Grundgesamtheit) zu erheben. Die Vorschriften hierzu liefern die im Rahmen der *Statistik* entwickelten Operationen, die vernünftige Entscheidungen auch noch in Fällen von Ungewißheit zu treffen erlauben.

**1.2.** Diese Ungewißheit, die im Kontext linguistischer Untersuchungen immer mehrere Gründe haben wird, charakterisiert Allén (1977) so:

> A corpus designed as a basis for the investigation of a language is of necessity a selection from the sum of manifestations. Given that it is possible to obtain such a corpus, it will as a sample tell us something about the sum of manifestations (the population), with the (un)certainty inherent in this kind of procedure. It will tell us something about the language system only indirectly. There is also the risk of errors, anacolutha, etc. in the corpus. Thus, in terms of information theory, there are two types of noise: statistical and linguistic/communicative (S. 1).

Was hier zunächst als quasi doppelte Störanfälligkeit der korpusgestützten linguistischen Forschung beschrieben wird, kann aber umgekehrt gerade innerhalb jener nur methodologisch faßbaren Wechselwirkung nutzbar gemacht werden, die bei allen empirisch-statistischen Ansätzen zwischen den *methodisch-operativen*, d. h. von der wahrscheinlichkeitstheoretischen Modellbildung der Inferenzstatistik bestimmten Bedingungen (statistical noise) einerseits und den *thematisch-inhaltlichen*, d. h. von der einbettenden Theorienbildung der (hier: linguistischen) Basiswissenschaft vorgegebenen Abhängigkeiten (linguistic/communicative noise) andererseits besteht. Dieser Interdependenz entspräche es dann auch, die genannten Bereiche der *Performanzanalyse*, der *Korpusbildung* und des *Computereinsatzes* als zwar unterscheidbare, einander aber wechselseitig

bestimmende und ergänzende Aspekte *eines* methodologischen Problemzusammenhangs zu begreifen, den ein empirisches Erkenntnismodell und die ihm zuordenbaren statistischen Methoden innerhalb der Linguistik zugleich herstellen, durchschaubar machen und nutzen lassen könnte.

**2.** Dem steht entgegen, daß die Rolle, welche der *Statistik* im Forschungsprozeß empirisch arbeitender Wissenschaften zukommt, in der Linguistik weder von ihren frühen Verfechtern noch auch von ihren späten Kritikern recht hat eingeschätzt werden können. Dies mag oberflächlich betrachtet zum Teil auf die vielleicht nur im Distributionalismus Harris'scher Prägung versuchte – heute bestenfalls noch im phonetisch-phonologischen Bereich akzeptierte – Anwendung strikt empirisch-quantitativer, Erkenntnismodelle zurückgehen; dies mag zum Teil auch mit der von Vertretern formaler (generativ-transformationeller) Grammatiktheorien vollzogenen Wende zur Kompetenz-orientierten Linguistik begründet werden, deren Empiriedistanz einer eigenständigen linguistischen Rezeption der Statistik nicht förderlich sein konnte. Aber es sollte doch nicht übersehen werden, daß selbst die jüngste über die Sozialwissenschaften und deren Rezeption vermittelte Auseinandersetzung mit der *Statistik* in der Linguistik untergründig noch von den frühen Einwendungen betroffen scheint, die schon gegen die Korpus-abhängig arbeitenden ‚discovery procedures' erhoben wurden. Die Pauschalierung dieser Kritik über die gesamte empirisch-quantitativ arbeitende Linguistik verhindert bis heute nicht nur, den schon angedeuteten Zusammenhang zwischen Veränderungen und Verschiebungen herzustellen, die sich nur zum Teil isoliert oder unabhängig voneinander in der *Performanzanalyse* (von Satz zum Text), in der *Korpusproblematik* (vom Umfang zur Repräsentativität) und im *Computereinsatz* (von automatischer Sprachanalyse zu künstlicher Intelligenz) inzwischen vollzogen haben, sondern sie schränkt gerade auch die Möglichkeiten ein, die eine vorurteilsfreie Rezeption empirisch-statistischer Modelle und ihrer Anendung in der linguistischen Forschung zu bieten vermöchte.

**2.1.** Die Ablehnung des vom taxonomischen Strukturalismus verfolgten empirischen Ansatzes war angesichts der deskriptiv-statistischen Methoden verständlich, die er auf ein Sprachkorpus anwandte, das primär als Belegsammlung distributionell bestimmter linguistischer Einheiten diente. Denn jedes Korpus solcher performatorischer Sprachdaten mußte – wie umfangreich auch immer angesetzt – jedenfalls endlich und damit unzureichend erscheinen, um aus ihm die Menge jener Strukturbeschreibungen erfolgreich zu ermitteln, die als Grammatik die Menge der akzeptablen Sätze in einer Sprache sollte beschreiben können.

**2.2.** Mit diesem grammatiktheoretisch begründeten Einwand hatten sich scheinbar auch die Probleme der Korpusbildung und des Computereinsatzes für die linguistische Forschung erledigt. Beider Diskussion war ja von der Frage nach dem *Umfang* der sprachlichen Datenmenge beherrscht gewesen, solange man geglaubt hatte, allein durch Vergrößerung des (mit Hilfe elektronischer Rechenanlagen auch zunehmend leichter zu verarbeitenden) Sprachkorpus sicherstellen zu können, daß alle (auch seltenere) linguistische Erscheinungen mit ausreichenden Häufigkeiten belegbar sein würden. Die Auffassung, wonach ein Korpus nicht nur als eine im philologischen Sinne verstandene Belegsammlung

dienen, sondern vielmehr als eine Stichprobe aus einer unendlichen Grundgesamtheit gedeutet werden könne, die prinzipiell immer nur ausschnitthaft zugänglich sein wird, weil kein Korpus alle möglichen vergangenen und zukünftigen Erscheinungsformen sprachlicher Performanz umfassen kann, diese Auffassung (vgl. etwa Ziff 1960) blieb vorerst ohne Konsequenz. Sie hätte auch in dieser frühen Phase der auf phonetisch-morphologischer Ebene des Satzes arbeitenden Untersuchungen schon vollziehen müssen, was erst für die post-generative Phase einer Performanz-orientierten, an der kommunikativen Funktion der Sprache ansetzenden Forschung gefordert werden kann: Der Übergang von der *deskriptiven* zur *urteilenden Statistik* in der empirischen Linguistik.
So aber wurde durch die neue Kompetenz-orientierte Grammatiktheorie der Empiriebezug in der Linguistik zunächst subjektiviert. Sie ersetzte das Korpus tatsächlich geäußerter Sätze durch Urteile kompetenter Sprachteilhaber über die Akzeptabilität von ad hoc gebildeten Satzbeispielen. Diese im Hinblick auf den idealen Sprecher/Hörer in homogener Sprachgemeinschaft bestenfalls heuristische nicht aber empirische Überprüfungsmethode der ‚Beispiele und Gegenbeispiele' (Schnelle 1970) bleibt auf die Satzebene beschränkt. Sie versagt daher nicht erst zur Beurteilung sondern schon bei der adäquaten Erfassung sowohl über-satzmäßiger als auch Sprecher-übergreifender Phänomene, die heute etwa in den (individuell, situativ, sozial, historisch oder wie immer beeinflußten) sprachlichen Variabilitäten (Klein 1976) einerseits, in den Bedeutung konstituierenden Funktionszusammenhängen kommunikativer Zeichen- und Sprachverwendung (Rieger 1977 b) andererseits angesprochen werden.

**3.** So gesehen bildet nicht nur der Mangel von empirischen Rückbindungsmöglichkeiten kompetenz-linguistischer Konstruktionen einen Ansatzpunkt der Kritik; es ist mehr noch das Unvermögen dieser auf den Satz fixierten formalen Grammatikmodelle, sehr augenfällige, von gesellschaftlichen Strukturen bestimmte und diese bestimmende Eigenschaften zu erfassen, wie sie an Texten als dem unmittelbar zugänglichen Resultat der Verwendung von Sprache zum Zwecke der Verständigung durch wirkliche Sprecher/Hörer in tatsächlichen Kommunikationssituationen beobachtet werden können. Dabei zeichnen sich zwei scheinbar isolierte aber nicht völlig beziehungslose Entwicklungen ab, die diesen Mängeln einer ausschließlich Kompetenz-orientierten, formalen Linguistik auf unterschiedliche Weise Rechnung tragen: im Bereich des *Computereinsatzes* die künstliche Intelligenz und im Bereich der *Performanztheorie* die Sozio- und Psycholinguistik, welche beide im Bereich der *Korpusbildung* aufeinander rückwirken könnten.

**3.1.** Der *Computereinsatz* in der Linguistik war während ihrer distributionalistisch-deskriptiven Phase mit den zunehmend größeren Datenmengen in den Korpora zu einer Notwendigkeit geworden angesichts der immer umfänglicheren Vergleichs-, Sortier- und Zählverfahren, die mit der Segmentierung und Klassifizierung der zu belegenden linguistischen Einheiten verbunden waren. Für diese fast ausschließlich auf phonologisch-morphologischer Ebene operierende maschinelle Sprachanalyse ist die Bezeichnung ‚Computational Linguistics' (Hays 1967; Bott 1970) bzw. ‚Computerlinguistik' (Dietrich/Klein 1974) geläufig geworden, der sich in der Phase der generativ-transformationellen Grammatiktheorie überdies neue Möglichkeiten der Anwendung eröffneten. So konnten die in verschiedenen Grammatikmodellen formulierten Regelsysteme im Rechner praktisch

durchgespielt und auf ihre Konsistenz und Effektivität hin überprüft werden (Friedmann 1971), wenn insgesamt auch von dieser Möglichkeit in geringerem Maße Gebrauch gemacht wurde als die Entwicklung der zahlreichen linguistischen Grammatikmodelle vermuten läßt.
Der Computerlinguistik nicht zugerechnet werden dagegen neuere Ansätze, die anwendungsorientiert in einem zwar nicht performanzanalytischen sondern eher funktionssimulierenden Sinne genannt werden können. Da sie im weitesten Sinne semantische und auch pragmatische Aspekte in den Vordergrund ihrer Untersuchung stellen und das Instrument des elektronischen Rechners im Rahmen seiner systemtheoretischen und prozeßsimulierenden Möglichkeiten einsetzen, hat sich für sie die Bezeichnung ‚Computational Semantics' (Raphael/Robinson 1972; Charniak/Wilks 1976) durchgesetzt. Diese unter dem Sammelbegriff der ‚künstlichen Intelligenz' (KI) unternommenen Forschungen beschäftigen sich mit simulativer Repräsentation von sprachlichem und außersprachlichem Wissen, Modellentwicklung des natürlich-sprachlichen Verstehens, Folgerns und Schließens, Organisation von Datenstrukturen in Netzwerken oder Entwicklung und Erprobung von Frage-Antwort-Systemen. Ihrer relativen Distanz zu linguistischer Theorienbildung im allgemeinen und zu empirisch deskriptiven Ansätzen im besonderen entspricht es, daß sie von Primärdaten und Eingangsinformationen ausgehen, die für je begrenzte situative Handlungsrahmen und Aufgabenstellungen nicht aufgrund empirischer Untersuchungen ermittelt, sondern von den Forschern mehr oder weniger ad hoc zusammengestellt werden. Darin dem subjektivierten Empiriebezug der Kompetenz-orientierten Grammatiktheorie vergleichbar, werden auch diese in ‚concepts', ‚frames' und ‚scripts' heuristisch verfügbar gemachten Basisinformationen zunehmend durch Ergebnisse verändert werden, die im Rahmen empirischer Untersuchungen einer Performanz-orientierten linguistischen Forschung erwartet werden können, wie umgekehrt auch Ergebnisse aus der KI-Forschung auf die Problem- und Fragestellung linguistischer Untersuchungen zur Semantik und Pragmatik zurückwirken werden (Rieger 1977 a; Wahlster 1977).

**3.2.** Für die post-generative Phase der linguistischen Performanzanalyse ist eine verstärkte Hinwendung zu gesellschaftswissenschaftlichen Problemen und eine zunehmende Beeinflussung durch sozialwissenschaftliche Forschungsansätze charakteristisch. Dies entspricht einer gerade aus den Mängeln der Kompetenz-orientierten Forschung abgeleiteten Einsicht in die primär gesellschaftlich wirksame Funktion von Sprache als Kommunikationsmittel, die es in ihren die sozialen Wirklichkeiten determinierenden Leistungen für Individuen, Gruppen und Schichten zu erforschen gilt (Hymes 1972) bzw. als pragmalinguistische Sprechhandlungstheorie in einem umfassenden handlungstheoretischen Zusammenhang zu stellen erlaubt (Wunderlich 1972). Der Übergang von der Satz- zur Äußerungs- bzw. Textebene ist für sie ebenso kennzeichnend, wie ihre um soziale Funktionen erweiterten Semantiktheorien, die auch den Kontext historischer, geographischer, pragmatischer etc. Normenvariabilitäten einzubeziehen suchen.
Für die praktische Forschungsarbeit bleibt der sozialwissenschaftliche Einfluß jedoch weitgehend auf eher *inhaltlich-thematische* Aspekte beschränkt. Ähnlich wie die bloße Anwendung des Computers auf sprachwissenschaftliche Gegenstände zur Computerlinguistik werden konnte, führt auch hier oft allein schon die Aufnahme und Behandlung von durch Ethnotheorie, Soziologie und/oder Psychologie angeregten Fragestellungen zu

jenen Etikettierungen, die gleichsam als Spezialgebiete (Pragmalinguistik, Soziolinguistik, Psycholinguistik) ausgliedern, was in die Linguistik allgemein besser hätte integriert werden können. Die Notwendigkeit zur Auseinandersetzung gerade auch mit der *methodisch-operativen* Herausforderung, die vor allem – im Hinblick auf die Möglichkeiten der KI-Forschung aber nicht nur – von den Sozialwissenschaften und deren empirisch-statistischen Methoden ausgeht, konnte so nachhaltig verdrängt und aus der Linguistik weg in quasi spezialistische Teildisziplinen abgeschoben werden. Dort freilich stellen sich die alten Probleme – wenn auch erst ansatzweise und zudem noch weitgehend isoliert – dem Bearbeiter erneut.

> Im linguistischen Bereich spielt im höheren Grade die Statistik eine Rolle. Dabei ergibt sich – freilich teilweise unter erheblicher Verbesserung der theoretischen Reflexion – eine erneute Hinwendung zur Verwendung von Korpora. (Steger 1973, S. 246).

Wie gering aber diese „Verbesserung der theoretischen Reflexion“ in Wirklichkeit ist, kann gerade der Zentralbereich des schon wiederholt hervorgehoben methodologischen Zusammenhangs von *Performanzanalyse* und *Computereinsatz* verdeutlichen: das Problem der *Korpusbildung*.

**4.** Einzig im Bereich der *Korpusproblematik* wird – wenn überhaupt – die *methodisch-operative* Seite bisher gesehen und eine nähere Bestimmung dessen versucht, was in Anlehnung an die Humboldt'sche Sprechweise „weniger dürftig“ und „zweckmäßig gewählt“ im Hinblick auf jeweils nur ausschnitthaft zugängliche Sprachrealisationen heißt. Diese vornehmlich im Rahmen sozio- und psycholinguistischer, aber auch lexikologischer und lexikographischer Untersuchungen angestellten Überlegungen scheinen sich dabei in der Forderung nach *Repräsentativität* einig, die sie aus teils performanzanalytischer, teils computerlinguistischer Sicht erheben.

Anhand schon einiger weniger (repräsentativer?) Beispiele läßt sich Richtung und Spannweite dieser Äußerungen zur Korpusbildung und Auswertung illustrieren. Dabei wird deutlich werden, daß die Verwendung des Repräsentativitätsbegriffs mehr zu verdecken als zu erhellen geeignet ist, welche Bedingungen linguistische Korpora erfüllen müssen, wenn sie nicht nur wie bisher *deskriptivstatistisch* analysiert werden, sondern im Rahmen *inferenzstatistischen* Argumentierens wie Stichproben fungieren sollen, deren Analyse im Hinblick auf Grundgesamtheiten neue Hypothesen zu bilden (Schätzen) bzw. vorliegende Hypothesen zu prüfen (Testen) erlaubt.

**4.1.** Im allgemeinen linguistischen Kontext wird bezeichnenderweise zunächst (noch) gar nicht von der Repräsentativität eines Korpus gesprochen. So betont Ziff (1960) im Gegenteil gerade die Unausweichlichkeit, mit der weder die Adäquatheit solcher als Stichproben zu verstehender Korpora abgeschätzt, noch auch die daraus ableitbaren Aussagen über die Grundgesamtheit aller in einer Sprache möglichen Äußerungen anders, denn als bestenfalls nur wahrscheinliche, d. h. jederzeit revidierbare Hypothesenbildungen eingeschätzt werden können.

> That $E$ (an indefinitely large but at most a finite corpus of utterances) is at best a sample of $E^*$, (the set of utterances composed of all the utterances that have been or will be uttered in the language) and a sample whose *adequacy*[1]) it is virtually impossible to assess, indicates that any conclusions about a natural language on the basis of a set like $E$ are at best merely *probable*. But this fact is neither cause for alarm nor a basis for skepticism. Much may be said about a natural language on the basis of a set like F. It is true that further evidence about the language, when available, may lead us to revise some of our views. But the fact that our views may be subject to revision in no way indicates either that they are not likely to be correct or that they are not worth expressing. (S. 20)

Bei Lyons (1968), der den statistischen Zusammenhang von Stichprobe und Grundgesamtheit ebenfalls aufnimmt, wird der Begriff der Repräsentativität eines Korpus – neben dem des Umfangs – als jene Bedingung eingeführt, die es erst ermöglichen soll, den strukturbeschreibenden Regeln, welche die grammatikalische Akzeptabilität der im Korpus belegten Äußerungen (Sätze) erklären, Gültigkeit auch zuzumessen für die unendliche Menge aller möglichen akzeptablen Äußerungen.

> The number of potential utterances in any natural language is unlimited. Any given collection of utterances, however large, is but a ‚sample' of this unlimited set of potential utterances. If the sample is not only large, but *representative* of the totality of potential utterances, it will, *ex hypothesi*, manifest all the regularities of formation characteristic of the language as a whole. (...) If rules are established to account for the acceptability of any *representative* sample of utterances, the same rules will necessarily account for a much larger set of utterances not in the original corpus, unless the application of the rules is very severely, and ‚unnaturally', restricted. Moreover, if certain rules with particular properties are incorporated into the description, it will be capable of accounting for an infinite, but specified, set of acceptable utterances. (S. 139f)

Unter direktem Hinweis auf den vorauszusetzenden wahrscheinlichkeitstheoretischen Zusammenhang nimmt Heger (1970) dagegen kritischen Bezug auf die Kriterien der Belegbarkeit, der Häufigkeit und der Akzeptabilität. Sie werden als unzureichend erklärt, um den notwendigen Übergang von den empirisch beobachtbaren Sprachdaten (parole) über die Gesamtheit aller darin aktualisierten Typen ($\sum$-parole) zum abstrakten System der solcher Typenkonstitution zugrunde liegenden Strukturregeln (langue) unmittelbar zu vollziehen.

> Jede Häufigkeitsanalyse, die den Vollzug des Übergangs von *parole* zu *langue* anstrebt, präsentiert sich nicht in Form einer – praktisch unmöglichen – exhaustiven Auszählung, sondern als mit Hilfe der Wahrscheinlichkeitsrechnung begründete Extrapolation aus einer exhaustiv analysierten Vorkommensmenge auf die unbekannte Obermenge der $\sum$-*parole*. Die Frage nach der Berechtigung solcher Extrapolationen ist dadurch zu beantworten, daß man die entsprechenden Regeln

[1] Hervorhebungen in den zitierten Textstellen wurden vom Verfasser B.R. vorgenommen.

> der Wahrscheinlichkeitsrechnung auf die Daten von Untermengen exhaustiv bekannter Vorkommensmengen anwendet und die so erhaltenen Resultate mit den tatsächlichen Daten der bekannten Gesamtmenge konfrontiert. Ergibt eine solche Konfrontation keine signifikante Abweichung, so ist das Extrapolationsverfahren gerechtfertigt, im anderen Falle ist es als unbrauchbar erwiesen. Beide Ergebnisse liegen aus entsprechenden Untersuchungen vor. (...) Während das Extrapolationsverfahren im phonetisch-phonologischen Bereich weitgehend als gerechtfertigt angesehen werden kann, ist es im lexikalischen Bereich infolge der Abhängigkeit der Häufigkeiten von thematischen und stilistischen Faktoren nachweisbar nicht anzuwenden. (S. 25)

Das Problem solcher Extrapolationen besteht aber wohl nicht (wenigstens zunächst nicht) im (semiotischen) Status der einer Untersuchung zugrunde gelegten (phonetisch-phonologischen, lexikalischen, syntaktischen oder semantischen) sprachlichen Einheiten, sondern vielmehr darin, daß im sprach- und textwissenschaftlichen Bereich oft gerade jene „Daten von Untermengen exhaustiv bekannter Vorkommensmengen“ auf den höheren semiotischen Ebenen fehlen. Denn nur wenn die Verteilung von jeweils betrachteten sprachlichen Einheiten in einer (welcher?) Grundgesamtheit bekannt ist, aus der überdies das Korpus, dem die beobachteten Daten entstandenen, als eine zufällige Stichprobe entnommen wurde, kann sinnvoll, d. h. mit angebbarem Risiko innerhalb eines statistischen Modells in Form einer Wahrscheinlichkeitsverteilung von der Stichprobe auf die Grundgesamtheit geschlossen werden.
Die Interdependenz von meist fiktiven Grundgesamtheiten und Zufallsstichproben im statistischen Argumentationszusammenhang einerseits, von Ziel und Gegenstand einer Untersuchung im Prozeß empirischer Erkenntnisgewinnung andererseits bilden die Pole eines wechselseitig abbildbaren Zusammenhangs, den Rieger (1972) als ein methodisches Grundprinzip der quantitativ-statistisch arbeitenden Sprach- und Textwissenschaft hervorhebt.

> So stellt sich für die mengenorientierte Textwissenschaft die statistische Methodenfrage in jeweils doppelter Weise: a) welche *Untersuchungsziele* können über eine mengenorientiert-textstatistische Analyse an einem vorliegenden Untersuchungsgegenstand sinnvoll angegangen (und gelöst) werden? b) welcher *Untersuchungsgegenstand* muß zugänglich sein, wenn ein angestrebtes Untersuchungsziel über eine mengenorientiert-textstatistische Analyse sinnvoll soll angegangen (und gelöst) werden können? Daraus erhellt, daß selbst ein mit größtem Aufwand durchgeführtes textstatistisches Analyseverfahren solange textwissenschaftlich irrelevant bleibt, wie eine Abgrenzung und Definition der Menge der untersuchten Texte die Interpendenz von Erkenntnisziel und Erkenntnisgegenstand nicht reflektiert und auf den die statistische Methodik konstituierenden wechselseitigen Zusammenhang von Zufälligkeit einer *Stichprobe* und Fiktivität ihrer *Grundgesamtheit* abbildet. (S. 27)

Eine eingehende Reflexion dieser Anhängigkeit von Erkenntnisziel und Erkenntnisgegenstand empirisch-statistischer Untersuchungen scheinen die Überlegungen zu bieten, die Nikitopoulos (1974) zum Korpusproblem im Rahmen eines sozio-linguistisch bestimmten

Ansatzes anstellt. Von der Frage nach der Repräsentativität eines Korpus der geschriebenen deutschen Gegenwartssprache ausgehend, erörtert er die Möglichkeit und Notwendigkeit einer allgemeinen Korpustheorie, welche erst die Bedingungen der Möglichkeit und Gültigkeit repräsentativer Korpora im Einzelfalle festlegen könne.

> Eine theoretische Begründung eines Corpus der geschriebenen deutschen Gegenwartssprache kann nicht nach irgendwelchen partikularen Kriterien erfolgen; sie muß vielmehr die Frage nach der *Repräsentativität* dieses Corpus beantworten, d.h. aber auch, das Bezugssystem explizit angeben, das ein gegenständlich gezieltes Fragen nach der Repräsentativität einer strukturierten Auswahl von geschriebenen Texten erlaubt. Das bedeutet wiederum, daß partikulare Fragestellungen, wie z.B. Untersuchungen über den Gebrauch des Konjunktivs, den Gebrauch des Perfekts und Präteritums in direkter Rede oder des öffentlichen Sprachgebrauchs usw., die u. U. spezielle oder Teilcorpora erfordern, nur im Rahmen einer allgemeinen Corpustheorie, die die Frage der Repräsentativität befriedigend beantwortet, gelöst werden können. Denn für diese speziellen oder Teilcorpora können nur im Rückgriff auf den transzendentalen Rahmen einer allgemeinen Corpustheorie die Bedingungen der Möglichkeit und Gültigkeit ihrer *Repräsentativität* angegeben werden. (S. 32f).

Mit einer solchen Thematisierung von Repräsentativität mußte sich die Problemstellung zugleich vertiefen und verengen: Denn nun geht es vor allem um die *inhaltliche* Abhängigkeit, mit der das, was überhaupt erst als Beobachtungsdatum gelten kann, von der jeweiligen (hier: sozio-linguistischen) Theorienbildung der einbettenden Basiswissenschaft konstituiert wird. Die *operative* Abhängigkeit dagegen, die mit der Frage nach den wahrscheinlichkeitstheoretischen Bedingungen verbunden ist, wie derartig konstituierte potentielle Beobachtungsdaten methodisch in tatsächlichen Korpora so zu erheben sind, daß sie als zufällige Stichproben aus Grundgesamtheiten gelten können, über die anders Kenntnis zu gewinnen unmöglich ist, diese Abhängigkeit bleibt außerhalb einer Konzeption von Repräsentativität, die oben *thematisch-inhaltlich* genannt wurde. Die Frage nach den *methodisch-operativen* Bedingungen der Stichprobenerhebung wird damit zum Randproblem, das Nikitopoulos als „nunmehr eher statistisch-technischer Art“ offenbar vernachlässigen zu können glaubt, wenn er schließt:

> Die Aufstellung eines repräsentativen Corpus ist dann eine mehr technische Frage nach den Stichproben-theoretischen Bedingungen der Möglichkeit und Gültigkeit einer Auswahl im Hinblick auf die relevanten Eigenschaften einer Grundgesamtheit. (S. 41).

**4.2.** Eine Klärung gerade der operativen Abhängigkeiten, deren Umsetzung in die Praxis jeder möglichen Basiswissenschaft eines der schwierigsten Probleme der angewandten Statistik darstellt, ist auch für die korpusabhängig arbeitende linguistische Forschung aber wohl erst noch zu leisten. Das jedenfalls läßt sich den (teilweise gleichlautenden) Formulierungen besonders der neuesten Publikationen entnehmen, mit der diese vermeintlich „mehr technische Frage“ bisher eher verschwommen gestellt als schon klar beantwortet wird.

Im Vordergrund steht hier das Problem der *Repräsentativität* des Korpus. Zu dessen Klärung sind die Aussagen der *Statistik* über die Bildung *repräsentativer Stichproben* heranzuziehen, zumal ja dem Korpus bei einer induktiv-statistischen Prüfung der jeweiligen Hypothese die Funktion einer *Stichprobe* zukommt (...). Solange z. B. nicht u. a. Vorkommen und relative Anteile der Redekonstellationstypen in einer bestimmten Zielgruppe bekannt sind, ist es nicht möglich, ein für den Sprachkonsum dieser Zielgruppe *repräsentatives Korpus* aufzustellen, d. h. ein Korpus, in dem die einzelnen Textsorten entsprechend dem Vorkommen und relativen Anteil von Redekonstellationstypen in eben dieser Zielgruppe *anteilmäßig zutreffend* und *quantitativ ausreichend* vertreten sind (Schank 1973, S. 22f).

Ausgehend von der prämisse, daß grundsätzlich jede sprachliche erscheinung in hinreichender menge in einem korpus belegbar und quantifizierbar sei, wird zunächst die frage nach der *mindestgröße* eines korpus und damit verbunden die nach dessen *repräsentativität* gestellt. Die zweite fragestellung ist eine technische, nämlich die nach der art des verfügbarmachens für die wissenschaftliche Auswertung (Bausch 1975, S. 128).

Ein zentrales Problem jeder auf eine Korpusauswertung angewiesenen Linguistik ist der Grad der *Repräsentativität* der ein Korpus bildenden Auswahl aus der Gesamtheit sprachlich realisierter Äußerungen (...). Diese *Auswahl* muß dann *repräsentativ* sein, wenn über die bloße Aufzählung von Korpusbefunden oder die Angabe von Korpusbelegen zum Zwecke der Illustration hinaus Klassifizierungen und damit verbunden verallgemeinernde und erklärende Aussagen über den Objektbereich angestrebt werden (...). Es wäre die definierte Menge aller individuellen Sprachrealisationen dann *angemessen repräsentiert*, wenn entsprechend der Häufigkeit ihres festgestellten Vorkommens bzw. Gebrauchs die jeweiligen Teilmengen aufgrund bestimmter *statistischer Kriterien* durch entsprechend ausgewählte Textexemplare *adäquat* vertreten wären (Schaeder 1976, S. 359/361).

Somit stellt sich als zentrales Problem der Korpusbildung das Problem der *repräsentativen Auswahl*. In den empirischen Sozialwissenschaften wird dieser Fragenkreis unter den Begriff der *Stichprobenbildung* diskutiert. Die dort erarbeiteten Ergebnisse sind jedoch nicht ohne weiteres auf linguistische Untersuchungen übertragbar. Weiterhin bleibt für den Linguisten die Aufgabe, zu entscheiden, welche Faktoren in seinem Bereich für die Erzielung von *Repräsentativität* wichtig sind, d. h. welche Faktoren des Kommunikationsaktes vernachlässigt bzw. konstant gehalten bzw. variiert werden können (...). Ein *repräsentatives Korpus* der deutschen Gegenwartssprache müßte die *gesamte Sprachwirklichkeit der Gegenwart modellhaft abbilden* (Schank/Schoenthal 1976, S. 17f).

Es darf als unbestritten gelten, daß es unmöglich ist, ein Korpus zu erstellen, das gleichsam eine *Dokumentation einer lebendigen Sprache darstellt*. Man wird stets auf eine *Auswahl*, die nach Möglichkeit *repräsentativ* sein sollte, angewiesen sein (...). Der Umfang eines Textkorpus, das der Lexikographie das Material für

> die Beschreibung liefern soll, ist auszurichten an der Forderung des *ausreichenden* Vorkommens der zu beschreibenden Lexeme in ihren je verschiedenen Verwendungsweisen (Bergenholtz/Schaeder 1977, S. 9).
>
> So entbrennt beispielsweise oft die wissenschaftliche Auseinandersetzung über die Frage nach der *Repräsentativität* eines erhobenen Corpus, nachdem der Einfluß der sozialwissenschaftlichen *Statistik* das Sammeln von ‚Belegen' (wie in der klassischen Dialektologie) als ungenügend erscheinen ließ. Dem Postulat der *repräsentativen Stichprobe* stehen in der Praxis aus technischen, personellen, finanziellen Gründen jedoch meist *nicht-repräsentative Zufallsstichproben* gegenüber, die probabilistische Aussagen über das Sprachverhalten einer anvisierten Population erlauben (Hess-Lüttich 1977, S. 14f).

Das alte Problem eines nach „Umfang" oder „Mindestgröße" durch „quantitativ ausreichendes", in „hinreichender Menge" oder durch „ausreichendes Vorkommen" der „sprachlichen Erscheinungen", „Textsorten", „Lexeme" oder kurz „Belege" bestimmten Korpus wird dabei durchaus gesehen. Und es wird ebenso auch – bei aller Unterschiedlichkeit der Bewertung – die neue Beziehung hervorgehoben, wonach die Korpusbildung durch Rückgriff auf die „Statistik" und deren „repräsentative Stichproben" aufgrund „bestimmter statistischer Kriterien" oder durch die „unter dem Begriff der Stichprobenbildung diskutierten" Anweisungen entweder *so* (Schank 73; Schaeder 76) oder *so* „nicht ohne weiteres" (Schank/Schoenthal 76) vorgenommen werden können. Deutlich uneinheitlicher wird aber schon die Möglichkeit eingeschätzt, ob Korpora als modellhafte Abbildungen oder Dokumentationen von Sprache fungieren können. Danach scheint eine „anteilmäßig zutreffende", „adäquate" oder „repräsentative Auswahl" zugleich *möglich* (Schank/Schoenthal 76), welche die „gesamte Sprachwirklichkeit der Gegenwart modellhaft abbildet" und *unmöglich* (Schank 73; Bergenholtz/Schaeder 77), wenn sie „gleichsam eine Dokumentation einer lebendigen Sprache darstellt."
Was endlich den durchgängig verwendeten Repräsentativitätsbegriff betrifft, so stellt er bestenfalls einen Namen dar für das „zentrale Problem" der Korpusbildung, aber gewiß nicht seine Lösung.

> Es ist wohl unmittelbar einleuchtend, daß von Stichproben auf Grundgesamtheiten nur sinnvoll geschlossen werden kann, wenn diese Stichproben die entsprechenden Grundgesamtheiten in allen wesentlichen Merkmalen möglichst gut widerspiegeln, d. h. wenn sie *repräsentativ* sind (...). Was das heißt, läßt sich leicht exakt *definieren*: Für jedes Element der Grundgesamtheit muß die gleiche Chance (Wahrscheinlichkeit) bestehen, in die (Zufalls-)Stichprobe aufgenommen zu werden. Die Forderung in die Tat umzusetzen, ist eines der schwierigsten Probleme der empirischen Forschung (Kriz 1973 S. 105f).

**5.** Wie schwierig es ist, diese Forderung zu verwirklichen und welche Fehlinterpretationen gerade der Repräsentativitätsbegriff nahelegt, zeigt die Gegenüberstellung von sogenannten „repräsentativen Stichproben" und „Zufallsstichproben" (Hess-Lüttich 77), die sich mit fast gleichen Worten überdies schon bei Schank (1973) findet:

> Ein aus einer *repräsentativen Stichprobe* aus einer anvisierten Population hervorgegangenes Korpus wäre eine optimale Materialgrundlage für linguistische Analysen. In der Regel gestalten sich jedoch Arbeitsaufwand und Kosten so aufwendig, daß man sich mit *nicht-repräsentativen Zufallsstichproben* begnügen muß, mit deren Hilfe Wahrscheinlichkeitsaussagen über das Vorkommen und die relativen Anteile von Redekonstellationstypen und Textsorten für eine Zielgruppe erarbeitet werden können (S. 23 Fn 12).

Denn erst die Unschärfe des Repräsentativitätsbegriffs, welcher die *Eigenschaften der Daten* in einem Korpus/einer Stichprobe einerseits von den *Eigenschaften des Verfahrens* zur Bildung von Korpora/Stichproben andererseits nicht zu unterscheiden erlaubt, kann jene verfehlte *Alternative* von entweder „repräsentativen“ oder „zufälligen“ Stichproben im Zusammenhang der linguistischen Korpusbildung sugerieren, der im Zusammenhang bestimmter Anwendungen von Statistik umgekehrt eine ebenso irrige *Identifikation* von „Zufallsstichproben“ mit „repräsentativen Stichproben“ entspricht.

> Ein weiteres Mißverständnis besteht darin, daß man eine Zufalls-Stichprobe unbedingt für eine „repräsentative“ Stichprobe oder einen „echten Querschnitt“ hält. Nur wenn man soviel über die Grundgesamtheit weiß, daß eine Stichprobe gar nicht notwendig ist, kann man garantieren, daß jede Stichproben-Methode, ob zufällig oder nicht, eine „repräsentative“ sein wird (...). Wo immer der Ausdruck „repräsentativ“ verwendet wird, um eine Stichprobe zu beschreiben, sollte man sehr sorgfältig untersuchen, was damit gemeint sein soll. Es ist unmöglich, die Auswahl einer Stichprobe zu garantieren, die repräsentativ für die Grundgesamtheit in Hinsicht auf Merkmale sein wird, die wir vor der Ziehung nicht kennen (Wallis/Roberts 1959, S. 280).

Wiederholte Hinweise auf oder Forderungen von *Repräsentativität* sollten daher nicht übersehen machen, daß dieser Begriff, der einer Stichprobe Zufälligkeit zugleich zu- und abzusprechen scheint, hinsichtlich dieses *methodisch-operativ* entscheidenden Kriteriums ambig bleibt. Seine im inferenzstatistischen Zusammenhang zudem nur zirkelhaft mögliche Verwendung läßt ihn daher nicht geeignet erscheinen, das Problem der Korpusbildung durchsichtiger zu machen.

**5.1.** Fragt man nämlich genauer nach der Bedeutung von *Repräsentativität* aufgrund des Gebrauchs, der von diesem relationalen Begriff im Hinblick auf Korpora gemacht wird, so ergeben sich wenigstens zwei Schwierigkeiten bei der Anwendung:

a) Ein Korpus soll offenbar dann *repräsentativ* genannt werden können, wenn es die – wie immer abgegrenzte – Menge (Grundgesamtheit), aus der es eine – wie immer bestimmte – Auswahl (Stichprobe) darstellt, hinsichtlich bestimmter Merkmale (Parameter) „entsprechend der Häufigkeit ihres festgestellten Vorkommens“, „anteilmäßig zutreffend“, „quantitativ ausreichend“ widergibt.
Sei etwa $A$ ein betrachtetes dichotomes Merkmal, sei die Wahrscheinlichkeit dafür, daß $A$ auftritt $P(A) = p$, und sei die relative Häufigkeit, mit der $A$ in einer Stichprobe vom Umfang $n$ $k$-mal beobachtet wird $k/n = f_n(A)$, so könnte diese Stichprobe hinsichtlich $A$ *repräsentativ* heißen, wenn

$$p = f_n(A)$$

Eine solche Gleichheit von Wahrscheinlichkeit und relativer Häufigkeit ist freilich extrem unwahrscheinlich. Es wird daher auch nicht von einer genauen Übereinstimmung ausgegangen, sondern vorsichtiger nur von „ausreichender“, „adäquater“, „modellhafter“, „angemessener“ Wiedergabe der in der Grundgesamtheit vorliegenden Verhältnisse durch die Stichprobe gesprochen. Diese unscharfen Prädikate lassen dabei völlig unbestimmt, bis zu welcher Differenz zwischen $p$ und $f_n(A)$ ‚noch gerade eben' bzw. ab welcher Differenz ‚schon nicht mehr' einer Stichprobe, *Repräsentativität* soll zugesprochen werden dürfen.

b) Der Begriff – und das ist hier entscheidender – muß seine derartig unscharfe Bedeutung offenbar aus der (über eine Grundgesamtheit) notwendig schon vorausgesetzten Kenntnis dessen beziehen, was (in einer Stichprobe) überhaupt *repräsentiert* werden kann.
Sei $P(A) = p$ wiederum die Wahrscheinlichkeit dafür, daß $A$ eintritt und sei $f_n(A)$ die relative Häufigkeit, mit der $A$ in einer Stichprobe vom Umfang $n$ beobachtet wird, so muß zur Feststellung einer mehr oder weniger guten Übereinstimmung beider Werte nicht nur $f(A)$, sondern unabhängig davon auch $p$ entweder als bekannt oder doch als ermittelbar gelten, wenn anders die Redeweise von der ‚Repräsentativität' einer Stichprobe nicht leer sein soll. Eben dies ist aber das Resultat, wenn man die Frage danach zu beantworten sucht, wie $P(A)$ definiert ist und möglicherweise berechnet werden kann.

Nach der unter Praktikern und Anwendern geläufigen Auffassung (R. v. Mises') kann die Wahrscheinlichkeit $P(A)$ als Grenzwert der relativen Häufigkeit $f_n(A)$ definiert werden, die mit wachsendem Stichprobenumfang $n$ gegen $P(A)$ strebt.

$$\lim_{n\to\infty} f_n(A) = P(A)$$

Diese Definition, wonach die Abweichung zwischen der beobachteten relativen Häufigkeit und der Wahrscheinlichkeit um so geringer sein wird, je größer die Zahl der gemachten Beobachtungen ist, muß als eine praktisch nicht einholbare Anweisung gelten. Obwohl sie bei der Diskussion der Frage nach dem *Umfang* von Korpora immer noch als ein ‚theoretisches' Fundament mißdeutet wird, erweist sie sich als kritisch gerade auch im Hinblick auf die Frage nach der *Repräsentativität* von Korpora.
Aus der Grenzwertdefinition der Wahrscheinlichkeit ergibt sich nämlich die paradoxe Folgerung, daß eine Stichprobe hinsichtlich des betrachteten Merkmals nur dann als *repräsentativ* ausgezeichnet werden kann, wenn über die Grundgesamtheit, aus der sie stammt, so viel bekannt ist, daß es eben dieser Stichprobenbildung gar nicht mehr bedarf. Da nach dieser Definition die Kenntnisse über die Grundgesamtheit (d. h. hinsichtlich des betrachteten Merkmals: seine Wahrscheinlichkeit) aber nur über Stichprobenbildung zu gewinnen sind, erweist sich zunächst einmal jede Kennzeichnung von Korpora als *zirkulär*, die deren Repräsentativität behauptet oder fordert. Zum anderen macht die das ideale (unter exakt gleichen Bedingungen beliebig oft wiederholbare) Zufallsexperiment voraussetzende Grenzwertdefinition der Wahrscheinlichkeit aber auch deutlich,

daß die Möglichkeit von Wahrscheinlichkeitsaussagen über und die Notwendigkeit der Zufälligkeitsforderung an tatsächliche Ereignisfolgen einen zwar in seinen Resultaten erfahrungsgemäß zutreffenden, selbst aber nicht *zirkelfrei definierbaren* Zusammenhang darstellt.

> Eine Stichprobe vom Umfang $n$ gilt als Zufallsstichprobe, wenn sie in einem Verfahren gewonnen ist, das jeder denkbaren Kombination von $n$ Einheiten in der Grundgesamtheit die gleiche Chance bietet, die Stichprobe zu sein, die tatsächlich gezogen wird (...). Die Grundidee kann man so fassen. Aber mit dem Wort ‚Chance' hat sich der Begriff Wahrscheinlichkeit in unsere Definition eingeschlichen, so daß wir uns im Kreise bewegen, indem wir Zufälligkeit durch Wahrscheinlichkeit und dann wieder Wahrscheinlichkeit durch Zufälligkeit definieren (Wallis/Roberts 1959, S. 262).

**5.2.** Aus diesem Dilemma hat die heute von der mathematischen wie angewandten Statistik allgemein anerkannte Auffassung (A. N. Kolmogorofs) die Konsequenz gezogen, eine Definition der Wahrscheinlichkeit erst gar nicht mehr zu versuchen. Sie versteht Wahrscheinlichkeit eines (zufälligen) Ereignisses als eine nicht explizit definierbare Grundgröße, die allerdings bestimmten Axiomen genügt. Danach ist die Wahrscheinlichkeit dafür, daß ein Ereignis $A$ eintritt, eine $A$ zugeordnete Zahl $P(A)$, die (1. Axiom) zwischen $0 \leq P(A) \leq 1$ liegt, die (2. Axiom) den Wert $P(E) = 1$ für sichere Ereignisse $E$ annimmt, und die (3. Axiom) bei unabhängigen, zufälligen Ereignissen $A_1$, $A_2$, ..., $A_m$ die Wahrscheinlichkeit dafür, daß $A_1$ oder $A_2$ oder ... $A_m$ eintritt, als Summe der Einzelwahrscheinlichkeiten ausdrücken läßt:

$$P(A_1 \vee A_2 \vee \ldots \vee A_m) = \sum_{i=1}^{m} P(A_i)$$

Auf diesen drei Axiomen baut die allgemeine Theorie der Wahrscheinlichkeit auf, welche die (mathematische) Grundlage aller statistischen Modelle bildet. Diese stellen verschiedene Typen (diskreter bzw. stetiger) Wahrscheinlichkeitsverteilungen von Zufallsvariablen dar, die jenen Zusammenhang mathematisch formulieren, dessen Resultate auch in der empirischen Realität mit umso besseren Näherungen beobachtet werden können, je größer die Anzahl dieser Beobachtungen ist. Gemeint ist damit der zunächst ja überraschende Sachverhalt, daß sich das Verhalten bestimmter numerischer Größen dann als regelhaft (und damit als vorhersagbar) charakterisieren läßt, wenn diese Größen als Resultate bestimmter Prozesse deutbar sind, für die der Einfluß des *Zufalls* konstitutiv ist.

So lassen sich etwa trotz aller Abweichungen, die sich beim Ziehen von Stichproben aus einer Grundgesamtheit ‚rein zufällig' ergeben, doch bestimmte Arten solcher Zufallsabweichungen (Typen der Stichprobenvariabilität) von Stichprobenparametern wie Mittelwert, Varianz, etc. unterscheiden, die einzig von den Grundgesamtheiten abhängen.

> Dieses Wissen um den Typ kann man nur mit Hilfe der Gesetze der mathematischen Wahrscheinlichkeit erhalten, und diese Gesetze gelten ausschließlich für Zufalls-Stichproben. Somit erlauben nur Zufalls-Stichproben objektive Generalisierungen

> aus der Stichprobe hinsichtlich der Grundgesamtheit (...). Der Statistiker hängt mithin von der Tatsache ab, daß der Typ der Variabilität von Zufalls-Stichproben aus jeder Grundgesamtheit durch die mathematischen Gesetze der Wahrscheinlichkeit bestimmt werden kann. Er erkennt nicht nur die Stichproben-Variabilität an, sondern er nutzt sie aus (Wallis/Roberts 1959, S. 89).

Die Kenntnis solcher Typen von Wahrscheinlichkeitsverteilungen (Stichprobenverteilungen), die angeben, mit welcher Wahrscheinlichkeit welche Werte eines Parameters dann erwartet werden können, wenn einzig *zufällige* (nicht aber irgendwelche anderen) Einflüsse wirksam sind, ist die Bedingung der Möglichkeit letztlich allen statistischen Schließens. Denn erst aufgrund dieser Wahrscheinlichkeitsverteilungen läßt sich einmal der Bereich (das Mutungsintervall), angeben, innerhalb dessen ein bestimmter Parameter mit einer vorgegebenen Wahrscheinlichkeit zu erwarten ist, und zum anderen prüfen, ob ein bestimmter Wert eines Parameters mit einer vorgegebenen Wahrscheinlichkeit (Signifikanzniveau) noch als zufällig gelten kann. Während im ersten Fall von den gemachten Beobachtungen ausgegangen wird, um aus ihnen auf die Grundgesamtheit zu schließen (Schätzen), wird im zweiten Fall schon von einer bestimmten Vorstellung über die Grundgesamtheit ausgegangen, um die Haltbarkeit dieser Vorstellung anhand der Beobachtungen zu überprüfen (Testen).

> Alle Schlüsse von Stichproben auf Grundgesamtheiten sind Wahrscheinlichkeitsschlüsse. Ist nun die Verteilung von bestimmten Variablen in der Grundgesamtheit bekannt und handelt es sich um eine *Zufallsstichprobe*, so läßt sich die Genauigkeit angeben, mit der diese Schlüsse vollzogen werden. Es bedarf dazu natürlich eines statistischen Modells, d. h. nur das statistische Verteilungsmodell in Form einer Wahrscheinlichkeitsverteilung erlaubt solche Schlüsse – das muß immer wieder betont werden (Kriz 1973, S. 106).

**6.** So gesehen ist die Sprechweise von repräsentativen Stichproben bzw. Korpora nicht etwa nur deswegen ungeeignet, weil sie unbestimmt und/oder notwendig zirkulär wäre, sondern sie ist vor allem deswegen als verfehlt zu verwerfen, weil sie außerstande setzt, das im wahrscheinlichkeitstheoretischen Begründungszusammenhang statistischen Schließens und Argumentierens vorausgesetzte *Kriterium der Zufälligkeit* zu erkennen und für das Verfahren der Korpusbildung zu fordern.
Daß diese Zufälligkeit der Korpusbildung im Rahmen der Performanz-orientierten linguistischen Forschung keine unerfüllbare *methodisch-operative* Forderung zu bleiben braucht, beruht auf der *thematisch-inhaltlichen* Interdependenz von Ziel und Gegenstand einer Untersuchung, die sich – wie an anderer Stelle (Rieger 1972) ausgeführt – auf den Zusammenhang von Grundgesamtheit und Stichprobe abbilden läßt.
Denn da aus der Sicht der Statistik ein *Untersuchungsziel* identisch ist mit dem Vorhaben, intersubjektiv nachprüfbare Aussagen unter (angebbarem) Risiko eines möglichen Fehlers über eine Grundgesamtheit zu machen, aufgrund von daraus entnommenen *zufälligen Stichproben*, die den empirisch zugänglichen *Untersuchungsgegenstand* (Korpus) bilden, kann dessen Zufälligkeit nicht nur durch das tatsächliche Verfahren seiner Erhebung im Hinblick auf ein vorgegebenes Untersuchungsziel sichergestellt, sondern auch dadurch erreicht werden, daß ein vorliegendes Korpus im Hinblick auf bestimmte,

allerdings dadurch eingeschränkte Untersuchungsziele als zufällig entnommen zumindest gedeutet werden kann.
Die Darlegung dieser und der nun tatsächlich „technischen“ Fragen der *methodisch-operativen* Anwendung statistischer Modelle muß aber einer eingehenderen Darstellung vorbehalten bleiben, welche unterschiedliche Problemstellungen aus dem linguistischen Bereich und deren Lösungsansätze im Zusammenhang von Performanzanalyse, Korpusbildung und Computereinsatz anhand von Beispielen vorführen und erläutern kann.

## Literatur


*Allén, Sture* (1977): Notes on the Role and Organization of a Corpus in the Investigation of a Language. Paper beim International Meeting of the Association for Literary and Linguistic Computing (ALLC) in Lüttich (Belgien).

*Bausch, Karl-Heinz* (1975): Zur Problematik der empirischen Basis in der Linguistik. Diskutiert am Modusgebrauch in Konditionalsätzen. ZGL 3. 1975, 123-148.

*Bellert, Irena* (1970). On a Condition of the Coherence of Texts. Semiotica 2. 1970, 335-363.

*Bergenholtz, H./Schaeder, B.* (1977): Deskriptive Lexikographie. ZGL 5. 1977, 2-33.

*Bott, M. F.* (1970): Computational Linguistics. In: Lyons, J. (Hrsg.): New Horizons in Linguistics. Harmondsworth 1970, 215-228.

*Brinker, Klaus* (1971): Aufgaben und Methoden der Textlinguistik. Kritischer Überblick über den Forschungsstand einer neuer linguistischen Teildisziplin. Wirkendes Wort 4. 1971, 217-37.

*Charniak, E/Wilks, Y.* (1976): Computational Semantics. An Introduction to Artificial Intelligence and Natural Language Comprehension. Amsterdam.

*Dietrich, R/Klein, W.* (1974): Computerlinguistik. Eine Einführung. Stuttgart.

*Dijk, Teun A. van* (1972): Some Aspects of Textgrammars. A Study in Theoretieal Linguistics and Poetics. The Hague/Paris.

*Dressler, Wolfgang* (1972): Einführung in die Textlinguistik. Tübingen.

*Friedmann, Joyce* (1971): A Computer Model of Transfonnational Grammar. New York.

*Hays, David G.* (1967): Introduction to Computational Linguisties. New York.

*Hartmann, Peter* (1971): Texte als linguistisches Objekt. In: Stempel, W.D. (Hrsg): Beiträge zur Textlinguistik. München 1971, 9-29.

*Heger, Klaus* (1970): Belegbarkeit, Akzeptabilität und Häufigkeit. Zur Aufgabenstellung der Sprachwissenschaft. In: Pilch, H./Richter, H. (Hrsg.): Theorie und Empirie in der Sprachforschung. Basel/München/New York 1970, 23-33.

*Hess-Lüttich, Ernest W. B.* (1977): Soziolinguistik und Empirie. Probleme der Corpusgewinnung und -auswertung: eine Einführung. In: Bielefeld, H.U./Hess-Lüttich, E.W.B./ Lundt, A. (Hrsg.): Soziolinguistik und Empirie. Beiträge zu Problemen der Corpusgewinnung und -Auswertung. Wiesbaden 1977, 10-28.

*Humboldt, Wilhelm von* (1829): Überr die Verschiedenheit des menschlichen Sprachbaues. In: Schriften zur Sprachphilosophie. Darmstadt 1963, 144-367.

*Hymes, D.* (1972): Editorial Introduction to Language in Society. In: Language in Society 1. 1972, 1-14.

*Klein, Wolfgang* (1976): Sprachliche Variation. Studium Linguistik 1. 1976, 29-46.

*Klein, Wolfgang* (1976): Einige wesentliche Eigenschaften natürlicher Sprachen und ihre Bedeutung für die linguistische Theorie. LiLi 23/24. 1976, 11-31.

*Kriz, Jürgen* (1973): Statistik in den Sozialwissenschaften. Einführung und kritische Diskussion. Reinbek.

*Lyons, John* (1968): Introduction to Theoretical Linguistics. London.

*Neurath, Paul* (1974): Grundbegriffe und Rechenmethoden der Statistik für Soziologen. In: Grundlegende Methoden und Techniken der empirischen Sozialforschung, 111. Teil. Stuttgart 1974.

*Nikitopoulos, Pantelis* (1974): Vorgriffe auf eine Thematisierung der Repräsentativität eines Corpus. deutsche sprache 1. 1974, 3 242.

*Petöft, Janos S.* (1972): Zu einer grammatischen Theorie sprachlicher Texte. LiLi 5. 1972, 31-58.

*Raphael, B./Robinson, A. E.* (1972): Bibliography on Computer Semantics. Stanford Research Inst. AI-Center, TN 7 2, 1972.

*Rieger, Burghard* (1972): Warum mengenorientierte Textwissenschaft? Zur Begründung der Statistik als Methode. LiLi 9. 1972, 11-28.

*Rieger, Burghard* (1977) a): Coling 76. Concepts, Frames, and Scripts in Aid of Semantic Networks, Knowledge Systems and Fantasies. SDv 1. 1977, 84-86.

*Rieger, Burghard* (1977 b): Bedeutungskonstitution. Einige Bemerkungen zur semiotischen Problematik eines linguistischen Problems. LiLi 27/28. 1977, 55-68.

*Schaeder, Burkhard* (1976): Maschinenlesbare Textkorpora des Deutschen und des Englischen. deutsche sprache 4. 1976, 356-370.

*Schank, Gerd* (1973): Zur Korpusfrage in der Linguistik. deutsche sprache 4. 1973, 16-26.

*Schank, G./Schoenthal, G.* (1976): Gesprochene Sprache. Eine Einführung in Forschungsansätze und Analysemethoden. Tübingen.

*Schnelle, H.* (1970): Theorie und Empirie in der Sprachwissenschaft. In: Pilch, H./ Richter, H. (Hrsg.): Theorie und Empirie in der Sprachforschung. Basel 1970, 51-65.

*Steger, Hugo* (1970): Einleitung. In: Steger, H. (Hrsg.): Vorschläge für eine strukturale Grammatik des Deutschen. Darmstadt 1970, Vll-XXII.

*Steger, Hugo* (1973): Soziolinguistik. In: LGL, hrsg. v. Altmann/Henne/Wiegand. Tübingen 1973, 245-254.

*Wallis, W. A./Roberts, H. V.* (1959): Methoden der Statistik. Ein neuer Weg zu ihrem Verständnis. Freiburg.

*Wahlster, Wolfgang* (1977): Die Repräsentation von vagem Wissen in natürlichsprachlichen Systemen (NSS) der künstlichen Intelligenz (K1). Bericht If1-HH-B-38/77 der Uni Hamburg.

*Wunderlich, D.* (197 1): Pragmatik, Sprechsituation, Deixis. LiLi 1/2. 1971, 153-190.

*Wunderlich, D.* (1972): Mannheimer Notizen zur Pragmatik. In: Maas, U./Wunderlich, D. (Hrsg.): Pragmatik und sprachliches Handeln. Frankfurt 1972, 279-294.

*Ziff, Paul* (1 960): Semantic Analysis. Ithaca, N. J.