Statistische Methoden in der Textlinguistik

Schmitz, Ulrich

In: Antos, Gerd; Brinker, Klaus; Heinemann, Wolfgang; Sager, Sven F. (Hg.): Text- und Gesprächslinguistik. Linguistics of Text and Conversation. Ein internationales Handbuch zeitgenössischer Forschung. 1. Halbband: Textlinguistik. Berlin, New York: de Gruyter

1. Motive der Textstatistik

"Wo Sinn ist", meint Wittgenstein (1960, 339 = SS 98), "muß vollkommene Ordnung sein", auch noch "im vagsten Satze". Das könnte man dann auch für den Signifikanten annehmen. Daher rührt das linguistische Motiv der Textstatistik: gibt es quantitative Eigenschaften der Ordnung von Texten? Ein zweites, verwandtes Motiv sucht Anwendungsbereiche der Mathematik jenseits der Natur in Erzeugnissen menschlichen Geistes. Ein drittes Motiv dient konkreten Anwendungen wie Autorenerkennung, Stilanalysen, Textoptimierung (z.B. durch Verständlichkeitsmessung) und Fremdsprachenlernen (z.B. durch Grundwortschatzbestimmung und Textauswahl). Alle drei Beweggründe setzen auch Untersuchungen über statistische Eigenschaften von Sprache als System und Sprachen als Systemen in Gang. Die folgende Darstellung gilt aber nicht der Sprachstatistik in diesem Sinne (vgl. Scholfield 1991), sondern der statistischen Untersuchung einzelner Texte und Textcorpora, auch wenn die statistischen Verfahren (Altmann 1995a, Gordesch 1991, Rietveld/van Hout 1993; Kurzdarstellungen Kauffer 1994, Schlobinski 1996, 87-167) und viele mathematische Konzepte (Piotrowski et al. 1985, 1990) grundsätzlich die gleichen sind. (Ein kompaktes Handbuch der quantitativen Linguistik liefert Tesitelová 1992, eine umfassende Bibliographie Köhler 1995.)

2. Aufgaben der Textstatistik

Textstatistik untersucht alle quantifizierbaren Eigenschaften von Texten, um sie zu charakterisieren, untereinander zu vergleichen und zu klassifizieren, auf historische, geographische, soziale oder psychologische Entstehungsbedingungen zu schließen und um Gesetze zu entdecken, die die Konstruktion von Texten steuern.

Sie beginnt mit der Definition und Zählung quantifizierbarer Einheiten von Texten. Solche rein deskriptiven Verfahren führen zu Häufigkeitstabellen (insbesondere Häufigkeitswörterbüchern) und statistischen Kenngrößen wie Mittelwerten und Indizes (z.B. Busemanns (1925) Aktionsquotient als Verhältnis von Adjektivanzahl zu Verbanzahl). Darüber hinaus verfolgt sie analytische Ansprüche und sucht eine "verborgene Ordnung" (Arens 1965) in Texten: sie spürt Wiederholungen (Altmann 1988) und überhaupt Mustern und Gleichförmigkeiten im Auftreten exakt definierter sprachlicher Einheiten nach. Wenn "speech is a series of nearly impossible events" (Geffroy et al. 1973, 129), so untersucht Textstatistik Wahrscheinlichkeiten in der Konstruktion von Texten in der Annahme, daß kommunikationstheoretische, anthropologische, psychologische, syntaktische, semantische und/oder pragmatische Gründe Abweichungen von völliger Zufallsverteilung erzwingen, und sei es nur um einer praktikablen Erzeugung und zugleich Reduktion von Komplexität willen.

Dabei geht es um die Erfassung von Trends, Tendenzen, Häufigkeitsverteilungen, stochastischen Abhängigkeiten, Korrelationen zwischen verschiedenen textinternen und textexternen Variablen und möglicherweise universellen Gesetzmäßigkeiten.

Mit Hilfe deskriptiver statistischer Verfahren werden also quantitative Eigenschaften von Texten bestimmt. Analytische Methoden bauen darauf auf und dienen dazu, Zusammenspiel, Konkurrenz und Entwicklung mehrerer Faktoren bzw. Merkmale beim Zustandekommen von Texten zu beschreiben und sprachliche Erzeugnisse auch als Ergebnis selbstregulierender Schemata und Prozesse zu verstehen (programmatisch Hrebicek/Altmann 1993) . Textstatistik insgesamt (1) zählt Textelemente aus und errechnet statistische Kennwerte von Texten, (2) mißt syntaktische und lexikalische Homogenität einzelner Texte oder einer Gruppe von Texten, (3) identifiziert Brüche innerhalb von Texten (sei es aufgrund besonderer Kreativität, Themen- oder Textsortenwechsels, schlechten Stils oder der Beteiligung verschiedener Autoren), (4) vergleicht Texte hinsichtlich quantifizierbarer Eigenschaften (z.B. um Stile, Epochen, Autoren oder Textsorten zu unterscheiden), (5) beschreibt probabilistische Charakteristika von Sprachnormen sowie Abweichungen bzw. Merkmale sprachlicher Varietäten (z.B. Fachsprachen, Soziolekte), Idiolekte oder einzelner Texte, (6) mißt und vergleicht lexikalische Reichhaltigkeit von Texten (z.B. durch Bestimmung der Anzahl verschiedener Wörter im Verhältnis zur Gesamtzahl der Wörter (type-token-ratio)), (7) mißt Verständlichkeit von Texten, soweit diese quantifiziert werden kann (vgl. Ballstaedt/Mandl 1988, Hrebicek/Altmann (eds.) 1993, 215-252), (8) beschreibt die allmähliche Entfaltung neuer Information in Texten (vgl. Wildgen 1993), (9) untersucht die lineare Präsentation nicht-linearen Wissens in Texten und (10) sucht allgemeine Eigenschaften, Unterschiede und Gesetzmäßigkeiten in Klassen aller Art von Texten (z.B. mündlich vs. schriftlich, Nachricht vs. Kommentar, Epik vs. Dramatik, Mittelalter vs. Moderne, Dialekt vs. Hochsprache) sowie (11) in "Text" überhaupt. (Einen gut verständlichen Querschnitt durch verschiedenartige Fragen und Methoden auf hohem Niveau bietet Tuldava 1995.)

3. Anwendungsbereiche in Beispielen

Textstatistische Verfahren können die Behandlung klassischer geisteswissenschaftlicher Gegenstände, soweit sie quantifizierbar sind, auf eine verläßliche empirische Grundlage stellen. Dazu zählen beispielsweise die Metrik (Grotjahn 1979) und die Entscheidung über die ggf. strittige Frage, von welchem oder welchen Autoren ein Text stammt (Wickmann 1989).

Statistik eröffnet aber auch neue, sonst nicht gestellte Fragen. Viele, vor allem die älteren, textstatistischen Arbeiten begnügen sich mit rein deskriptiven Verfahren, zählen also Elemente aus (z.B. die bei Harkin 1957 und Billmeier/Krallmann 1969 genannten, so etwa Krallmann 1966, Meier 1967) und erstellen etwa Häufigkeitswörterbücher (z.B. Ruoff 1981). Unerläßlich sind textstatistische Verfahren bei der Analyse und ggf. auch Konstruktion großer Textcorpora (vgl. Bergenholtz/Schaeder (ed.) 1979, Leech 1991, Stubbs 1996).

Oft werden sowohl einzelne Texte als auch ganze Textcorpora als samples für vermutete Gesetzmäßigkeiten im sprachlichen System (und teilweise auch allgemeineren Gegebenheiten) statistisch untersucht (z.B. Brainerd 1971, Grotjahn 1982, Herdan 1966, Schmidt (ed.) 1996). Das gilt insbesondere für das Zipfsche Gesetz (wegen des Grundprinzips des geringsten Kraftaufwandes ist das Produkt aus Häufigkeits-Rangplatz und Verwendungshäufigkeit von Wörtern in Texten stets konstant; Zipf 1932, 1935, Guiter/Arapov (eds.) 1982) und die Menzerathsche Hypothese (je größer ein sprachliches Ganzes, desto kleiner seine Teile; Menzerath 1954, Altmann/Schwibbe 1989, Hrebicek 1995).

Häufig verfolgt werden auch lexikographische (Menzerath 1954, Hellmann (ed.) 1984) und stilistische Fragestellungen (Überblick bei Hoffmann/Piotrowski 1979, 148-156; später Pieper 1979). Brainerds (1972) Untersuchung des Artikelgebrauchs als Stilindikator ist ein kleines, aber sehr typisches Beispiel.

Seltener, aber meist sehr ergiebig, sind analytisch-statistische Untersuchungen zur Eigenart einzelner Texte (Orlov u.a. 1982), zu semantischen Relationen in Texten (Skorochod'ko 1981, 120-185), zur dynamischen Entwicklung von Merkmalen im Verlauf eines Textes (z.B. Entropie und Wiederholungsrate) (Köhler/Galle 1993), zu Entwicklungslinien in der Schreibweise eines einzelnen Autors (Laffal 1997), zu langfristigen Entwicklungen im Vokabular und damit verbundenen spezifischen Einstellungsänderungen in der Bevölkerung (z.B. Fortier/Keen 1997). Und schließlich können statistische Textuntersuchungen auch dazu beitragen, die Leistungsfähigkeit von Programmen zur maschinellen Erzeugung oder Analyse natürlich-sprachlicher Texte zu verbessern (vgl. z.B. Walker/Moore 1997).

4. Statistische Methoden und wissenschaftliche Theoriebildung

Mit statistischen Methoden können nur quantifizierbare Eigenschaften von Texten erfaßt werden. "Information" beispielsweise als Maß für die Unwahrscheinlichkeit des Auftretens eines Elements kann gemessen werden, "Sinn" aber nicht. Damit ist die grundsätzliche Frage nach der besonderen Leistung menschlicher Sprache aufgeworfen. ("Die Form der Zahl und des Zählens ist daher das eigentliche Bindeglied, an welchem man sich den Zusammenhang zwischen sprachlichem und wissenschaftlichem Denken, wie den charakteristischen Gegensatz zwischen beiden am deutlichsten vergegenwärtigen kann." Cassirer 1953/1954, Bd.3, 399) Die Beziehungen zwischen quantitativen, symbolorientierten, strukturellen und hermeneutischen Zugangsweisen sind aufgrund gerne sich abkapselnder Schulbildungen noch nicht genügend diskutiert worden (für die beiden erstgenannten vgl. Klavans/Resnik (ed.) 1996).

Während, um ein Beispiel zu nennen, die traditionelle Stilistik stark auf subjektive Urteilskraft baut, untersucht die quantitative Stilistik zähl- und also objektivierbare stilistische Merkmale. Ob und in welcher Weise beide Seiten voneinander profitieren können, ist kaum hinreichend konkret bedacht worden. "Der quantitative Ansatz vermag zwar aufzudecken, wie sich ein Einzelwerk oder auch eine verwandte Gruppe von Texten zu Sprach-, Textgruppen- oder auch Epochennormen verhält, die Interpretation der Übereinstimmung oder der Abweichung von diesen Normen in Richtung auf ein Versagen des Autors, die gesetzte Normierung zu erreichen oder eher in Richtung auf einen Erfolg, beispielsweise einen Innovationseffekt erzielt zu haben, wird Aufgabe der qualitativen Stilistik bleiben. Die quantitative Analyse schmälert also in keiner Weise eine traditionell ausgerichtete Literaturbetrachtung oder Literaturkritik. Sie liefert ihr vielmehr Werkzeug und Daten, um ihre qualitativen Aussagen empirisch zu belegen" (Pieper 1979, 125).

In der Regel führen diejenigen Untersuchungen am weitesten, die ihre statistischen Analysen aus einem größeren Reflexionszusammenhang begründen. Statistik ihrerseits zwingt zur Formulierung überprüfbarer Aussagen und wirkt dadurch disziplinierend, aber auch belebend auf wissenschaftliche Begriffsbildung und Methodik. Einerseits dient sie der Überprüfung vorab formulierter Hypothesen (z.B. über den Vergleich einzelner Texte oder Stichproben untereinander, über das Verhältnis von Stichprobe und Grundgesamtheit, über das Verhältnis

von beobachteten Daten und theoretischer Funktion oder Verteilung); und sie erlaubt die Vorhersage nicht beobachteter aufgrund von beobachteten Daten, die Überprüfung der Qualität einer Stichprobe sowie den Vergleich verschiedener Klassifikationen (z.B. von Textsorten) untereinander. Andererseits erfüllt sie aber auch eine heuristische Funktion und lädt zur Formulierung sonst vielleicht gar nicht erdachter Hypothesen ein, nämlich wenn (oft überraschende) Korrelationen zwischen Variablen aufgefunden werden (z.B. durch Faktorenanalyse oder pfadanalytische Verfahren).

Freilich stehen alle textstatistischen Untersuchungen vier Schwierigkeiten gegenüber.

(1) Trivialerweise hängt das Ergebnis der Arbeit stark von der Definition der untersuchten Texteinheiten ab. Für Phoneme und Buchstaben, für Silben und Morpheme, für Lemmata und Wortformen, für Syntagmen und Phrasen, für Sätze und Redeeinheiten (turns) gelten nicht unbedingt ähnliche Verteilungen oder Gesetzmäßigkeiten. Es ist aber nicht leicht, die untersuchten Texteinheiten exakt zu definieren. (Selbst bei der einfachsten Definition von ‚Wort' als Buchstabenfolge zwischen Leerräumen können unterschiedliche Zählungen zustande kommen. Für viele linguistische Kategorien, z.B. Wortarten, gibt es keine hinreichend genaue - intersubjektiv verläßliche - intensionale Definition.) Und noch schwerer ist es oft, eine präzise Definition zu finden, die auch für die Fragestellung taugt. (Beispielsweise sollten trennbare Verben in inhaltsorientierten Untersuchungen als ein Wort aufgefaßt werden. Oder vorgängige Textsortenunterscheidungen erfassen die charakteristische Zusammensetzung des untersuchten Corpus nicht.) Deshalb sind ähnliche Untersuchungen nicht ohne weiteres miteinander vergleichbar; und die Einzelfallbeschreibung kann nicht immer für einen größeren Bereich oder für allgemeinere Aussagen fruchtbar gemacht werden.

(2) Die untersuchten Textelemente (z.B. Wörter) - im Gegensatz zu einigen ihnen äußerlich zukommenden Eigenschaften (z.B. Wortlänge: Verhältnisskala; Position im Satz oder Text: Ordinalskala) - werden auf einer Nominalskala gemessen. Dafür können aber in der deskriptiven Statistik nur die am wenigsten informationshaltigen Parameter (Modus für die Lage, Häufigkeitsverteilung für die Streuung und Kontingenzkoeffizient für die Korrelation) und in der analytischen Statistik nur einige wenige Schätz- und Entscheidungsverfahren verwendet werden.

(3) Fast alle statistischen Verfahren und die meisten Modelle wurden im Rahmen natur- und sozialwissenschaftlicher Fragestellungen entwickelt. Sie passen nicht ohne weiteres zum Gegenstand Sprache und laden bei textlinguistischer Übertragung zu Fehlern ein. Verglichen mit anderen statistisch orientierten Wissenschaftszweigen steht Textstatistik erst am Anfang ihrer Entwicklung. Viele Untersuchungen orientieren sich an dem, was statistisch leicht möglich ist. Es ist nicht immer einfach, das angemessene Verfahren und das passende mathematische Modell für eine genuin textwissenschaftliche Fragestellung zu finden. "Which methods should be applied in order to grasp processes which present themselves as time series, stochastic and chaotic sequences? Does the text have its own mathematics that has not been discovered as yet?" (Altmann 1995b, V)

(4) Es ist nicht leicht, von der Beschreibung einer Reihe einzelner Merkmale zu einer erkenntnisträchtigen allgemeineren Charakterisierung des Textes oder Textcorpus bzw. von der Beschreibung eines einzelnen Textes oder Textcorpus zur Entdeckung allgemeiner Gesetze zu kommen. Beim derzeitigen Stand der Textstatistik stehen verfahrenstechnischer Aufwand und wissenschaftlicher (auch verallgemeinerbarer theoretischer) Ertrag oft nur in einem unbefriedigenden Verhältnis (vgl. Schmitz 1983).

All das spricht nicht etwa gegen, sondern für Einsatz und Weiterentwicklung statistischer Methoden in der Textlinguistik. Man muß sich nur ihrer teils aktuellen, teils prinzipiellen Grenzen bewußt sein.

5. Ausblick: Sprache und Text

In mathematischer Hinsicht können Texte als Ergebnisse stochastischer, dynamischer, nicht-rekursiver, nicht-stationärer, offener und zielsuchender Prozesse betrachtet werden (vgl. Altmann/Grotjahn 1988, 1026f; Hrebicek 1993). "Der Zusammenhang zwischen der strukturellen Unvollkommenheit des Systems ‚Sprache` und seiner Wandlungsfähigkeit zum Ausdruck aller möglichen Gedanken läßt sich erst im Rahmen der mathematischen Chaosforschung erkennen." (Bluhme 1988, 6) Wenn dies gelänge, könnte eine quantitativ orientierte Text- und Sprachbetrachtung dazu beitragen, die künstliche Unterscheidung von Regel und Anwendung, von System und Gebrauch zu überwinden und vielmehr "die Sprache" in der Gesamtheit "des jedesmaligen Sprechens" zu sehen (vgl. Humboldt 1963, 418). Auf diese Weise könnten textstatistische Untersuchungen auch helfen, sprachgeschichtliche Tendenzen (vgl. Embleton 1986) "als notwendige unbeabsichtigte Konsequenz individueller Handlungen auszuweisen, die unter bestimmten ökologischen Bedingungen nach bestimmten Handlungsmaximen vollzogen worden sind" (Keller 1990, 199).

Freilich bewährt sich Textstatistik nur im mühseligen Alltag handwerklich sorgfältiger Einzeluntersuchungen. Dabei sollte jeweils eine theoretisch wohldurchdachte Fragestellung Datenerhebung, -auswertung und -interpretation bis ins einzelne leiten. Sonst versinkt man in unübersichtlichen Zahlengräbern von geringem Erkenntniswert (z.B. Rohrmann 1974), weil auch bei noch so objektiven Verfahren "die Vernunft nur das einsieht, was sie selbst nach ihrem Entwurfe hervorbringt" (Kant 1956, 23 = B XIII).

6. Literatur (in Auswahl)

- Altmann, G[abriel] (1988): Wiederholungen in Texten. Bochum.
- Altmann, Gabriel (1995a): Statistik für Linguisten [1980]. Trier.
- Altmann, G[abriel] (1995b): Preface. In: Tuldava 1995, I-VII.
- Altmann, Gabriel/ Grotjahn, Rüdiger (1988): Linguistische Meßverfahren. In: Ammon, Ulrich/ Dittmar, Norbert/ Mattheier, Klaus J. ed.: Sociolinguistics. Soziolinguistik. An International Handbook of the Science of Language and Society. 2. Halbband. Berlin/ New York, 1026-1039.
- Altmann, Gabriel/ Schwibbe, Michael H. (1989): Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Hildesheim/ Zürich/ New York.
- Arens, Hans (1965): Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute. Düsseldorf.
- Ballstaedt, Steffen-Peter/ Mandl, Heinz (1988): The Assessment of Comprehensibility. In: Ammon, Ulrich/ Dittmar, Norbert/ Mattheier, Klaus J. ed. : Sociolinguistics. Soziolinguistik. An International Handbook of the Science of Language and Society. 2. Halbband/ Berlin/ New York, 1039-1052.
- Bergenholtz, Henning/ Schaeder, Burkhard. ed. (1979): Empirische Textwissenschaft. Aufbau und Auswertung von Textcorpora. Königstein/Ts.
- Billmeier, G./ Krallmann, D. (1969): Bibliographie zur statistischen Linguistik. Hamburg (Forschungsbericht 69/3 des Instituts für Kommunikationsforschung und Phonetik der Universität Bonn).

- Bluhme, Hermann (1988): Zur Einleitung: Linguistik ohne Maß und Zahl? In: ders. ed.: Beiträge zur quantitativen Linguistik. Gedächtniskolloquium für Eberhard Zwirner. Tübingen, 5-8.
- Brainerd, Barron (1971): Introduction to the mathematics of language study. New York.
- Brainerd, Barron (1972): Article use as an indicator of style among English-language authors. In: Jäger, Siegfried. ed.: Linguistik und Statistik. Braunschweig, 11-32.
- Busemann, Adolf (1925): Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik. Jena
- Cassirer, Ernst (1953/1954): Philosophie der symbolischen Formen [1923-1929]. 3 Bde. 2. Aufl. Darmstadt.
- Embleton, Sheila M. (1986): Statistics in historical linguistics. Bochum.
- Fortier, Paul A./ Keen, Kevin J. (1997): Change Points: Ageing and Content Words in a Large Database. In: Literary and Linguistic Computing 12, 14-22.
- Geffroy, Annie/ Lafon, P./ Seidel, Gill/ Tournier, M. (1973): Lexicometric analysis of co-occurrences. In: Aitken, A. J./ Bailey, R. W./ Hamilton-Smith, N. eds.: The Computer and Literary Studies. Edinburgh, 113-133.
- Gordesch, Johannes (1991): Statistische Datenverarbeitung in der Textanalyse. Berlin: Freie Universität, Institut für Soziologie.
- Grotjahn, Rüdiger (1979): Linguistische und statistische Methoden in Metrik und Textwissenschaft. Bochum.
- Grotjahn, Rüdiger (1982): Ein statistisches Modell für die Verteilung der Wortlänge. In: Zeitschrift für Sprachwissenschaft 1, 44-75.
- Guiter, H./ Arapov, M. V. eds. (1982): Studies on Zipf's Law. Bochum.
- Harkin, Duncan (1957): The History of Word Counts. In: Babel 3, 113-124.
- Hellmann, Manfred W. eds. (1984): Ost-West-Wortschatzvergleiche. Maschinell gestützte Untersuchungen zum Vokabular von Zeitungstexten aus der BRD und der DDR. Tübingen.
- Herdan, Gustav (1966): The Advanced Theory of Language as Choice and Chance. Berlin, Heidelberg, New York.
- Hoffmann, L./ Piotrowski, R. G. (1979): Beiträge zur Sprachstatistik. Leipzig.
- Hrebicek, L[udek] (1993): Text as a strategic process. In: Hrebicek, L[udek]/ Altmann, G[abriel] eds.: Quantitative Text Analysis. Trier, 136-150
- Hrebicek, L[udek] (1995): Text Levels. Language Constructs, Constituents, and the Menzerath-Altmann Law. Trier.
- Hrebicek , L[udek]/ Altmann, Gabriel (1993): Prospect of text linguistics. In: Hrebicek, L[udek]/ Altmann, G[abriel] eds.: Quantitative Text Analysis. Trier, 1-28.
- Hrebicek, L[udek]/ Altmann, G[abriel] eds. (1993): Quantitative Text Analysis. Trier.
- von Humboldt, Wilhelm (1963): Ueber die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts [1830-1835]. In: ders.: Werke in fünf Bänden (ed. Andres Flitner/ Klaus Giel), Bd. III: Schriften zur Sprachphilosophie. Darmstadt, 368-756.
- Kant, Immanuel (1956): Kritik der reinen Vernunft [1781]. (= Werke, ed. Wilhelm Weischedel, Bd. II). Wiesbaden.
- Kauffer, Maurice (1994): Le linguistique et la statistique. In: Nouveaux Cahiers d'allemand 12, no. 1, 55-91.
- Keller, Rudi (1990): Sprachwandel. Von der unsichtbaren Hand in der Sprache. Tübingen.
- Klavans, Judith/ Resnik, Philip. eds. (1996): The Balancing Act. Combining Symbolic and Statistical Approaches to Language. Cambridge, MA.

- Köhler, Reinhard (with the assistance of Christiane Hoffmann) (1995): Bibliography of Quantitative Linguistics (Bibliographie der quantitativen Linguistik; Bibliografija po kvantitativnoj lingvistike). Amsterdam/ Philadelphia.
- Köhler, Reinhard/ Galle, Matthias (1993): Dynamic aspects of text characteristics. In: Hrebicek, L[udek]/ Altmann, G[abriel]. eds.: Quantitative Text Analysis. Trier, 46-53.
- Krallmann, Dieter (1966): Statistische Methoden in der stilistischen Textanalyse. Phil. Diss. Bonn.
- Laffal, Julius (1997): Union and Separation in Edgar Allan Poe. In: Literary and Linguistic Computing 12, 1-13.
- Leech, Geoffrey (1991): The state of the art in corpus linguistics. In: Aijmer, Karin/ Altenberg, Bengt. eds.: English Corpus Linguistics. Studies in Honour of Jan Svartvik. London, 8-29.
- Meier, Helmut (1967): Deutsche Sprachstatistik. 2 Bde. [1964]. 2. Aufl. Hildesheim.
- Menzerath, Paul (1954): Die Architektonik des deutschen Wortschatzes. Bonn.
- Orlov, Ju. K./ Boroda, M. G./ Nadarejsvili, I. .. (1982): Sprache, Text, Kunst. Quantitative Analysen. Bochum.
- Pieper, Ursula (1979): Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse. Tübingen.
- Piotrowski, R. G./ Bektaev, K. B./ Piotrowskaja, A. A. (1985): Mathematische Linguistik. Bochum.
- Piotrowski, R./ Lesohin, M./ Lukjanenkov, K. (1990): Introduction of Elements of Mathematics to Linguistics. Bochum.
- Rietveld, Toni/ van Hout, Roeland (1993): Statistical Techniques for the Study of Language and Language Behavior. Berlin, New York.
- Rohrmann, Bernd (1974): Psychometrische und textstatistische Studien zu syntaktischen Variablen. Hamburg.
- Ruoff, Arno (1981): Häufigkeitswörterbuch gesprochener Sprache: gesondert nach Wortarten, alphabetisch, rückläufig alphabetisch und nach Häufigkeit geordnet. Tübingen.
- Schmidt, Peter. ed. (1996): Issues in General Linguistic Theory and The Theory of Word Length. Trier.
- Schmitz, Ulrich (1983): Zählen und Erzählen - Zur Anwendung statistischer Verfahren in der Textlinguistik. In: Zeitschrift für Sprachwissenschaft 2, 132-143.
- Schlobinski, Peter (1996): Empirische Sprachwissenschaft. Opladen.
- Scholfield, Phil (1991): Statistics in linguistics. In: Annual Review of Anthropology 20, 377-393.
- Skorochod'ko, E. F. (1981): Semantische Relationen in der Lexik und in Texten. Bochum.
- Stubbs, Michael (1996): Text and Corpus Analysis. Computer-assisted Studies of Language and Culture. Oxford.
- Tesitelová, Marie (1992): Quantitative linguistics. Amsterdam, Philadelphia: Benjamins
- Tuldava, Juhan (1995): Methods in Quantitative Linguistics. Trier.
- Walker, Marilyn A./ Moore, Johanna D. (1997): Empirical Studies in Discourse. In: Computational Linguistics 23, 1-12.
- Wickmann, Dieter (1989): Computergestützte Philologie: Bestimmung der Echtheit und Datierung von Texten. In: Bátori, István/ Lenders, Winfried/ Putschke, Wolfgang. eds.: Computational Linguistics. Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer [sic] Anwendungen. Berlin/ New York, 528-534.

- Wildgen, Wolfgang (1993): The distribution of imaginistic information in oral narratives. A model and its application to thematic continuity. In: Hrebicek, L[udek]/ Altmann, G[abriel]. eds.: Quantitative Text Analysis. Trier, 175-199.
- Wittgenstein, Ludwig (1960): Philosophische Untersuchungen [1953]. In: ders.: Schriften 1. Frankfurt/M., 279-544.
- Zipf, George Kingsley (1932): Selected Studies of the Principle of Relative Frequency in Language. Cambridge (Mass.).
- Zipf, George Kingsley (1935): The Psycho-Biology of Language. Cambridge (Mass.).