QUANTITATIVE LINGUISTICS

Volume 58

Karl-Heinz Best (ed.)

Glottometrika 16

The Distribution of Word and Sentence Length

Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Glottometrika ... -
WVT Wissenschaftlicher Verlag Trier.
Früher mehrbd. begrenztes Werk. -
Bis 13 (1992) im Verl. Brockmeyer, Bochum
ISSN 0932-7991
Bd. 16. The distribution of
of word and sentence length. - 1997
The distribution of word and sentence length / Karl-Heinz Best (ed.). -
WVT Wissenschaftlicher Verlag Trier, 1997
(Glottometrika ... ; 16)
(Quantitative linguistics ; Vol. 58)
ISBN 3-88476-276-1

Umschlag: Brigitta Disseldorf
(Marco Nottar, Agentur für Werbung und Design, Konz)

ISBN 3-88476-276-1
ISSN 0932-7991

WVT Wissenschaftlicher Verlag Trier
Bergstraße 27, 54295 Trier
Postfach 4005, 54230 Trier
Tel. 0651-41503, Fax 41504

Inhaltsverzeichnis

Vorwort

Wortlänge

Satzlänge

Wortarten

Glottometrika 16, 1997, v-xii

Warum nur: Wortlänge? Nicht nur ein Vorwort

Karl-Heinz Best

Wie häufig kommen Wörter und Sätze verschiedener Länge in einem vollständigen Text vor und welche Gesetzmäßigkeiten liegen den beobachteten Verteilungen zugrunde? Mit dieser Frage befassen sich die Arbeiten in dem vorliegenden Sammelband. Hinzu kommt ein Beitrag zur Wortartenhäufigkeit in deutschen Texten.

Obwohl Wörter und Sätze ja zweifellos zentrale Einheiten der Sprachstruktur sind und die Bedeutung der Frequenz von Einheiten seit langem bekannt ist, kann man mit kaum einer Themenstellung mehr Unverständnis auslösen als mit der genannten, wie Mitarbeiter des Projekts zur Wortlängenforschung (Best & Altmann, 1996) immer wieder berichten. Deshalb soll hier versucht werden, die Rolle von Wort- und Satzlängen zu skizzieren und ihren Zusammenhang mit der Frequenz aufzuzeigen, bevor auf die spezielle Thematik dieses Bandes und ihre Einbindung in die Theorie der Quantitativen Linguistik eingegangen wird. Dabei stehen Wortlängen entsprechend der Entwicklung des Projekts und seiner Schwerpunktbildung im Vordergrund.

Wenn man sich der Wortlänge zunächst aus nichtlinguistischer Sicht, also „von außen", nähert, kann man z.B. im Zusammenhang mit der Behandlung von Eigennamen auf einschlägige Hinweise stoßen. So hat schon Jean Paul (^{2}1813:270) bekannt, daß er „unbedeutende Menschen einsilbig: Wutz, Stuß getauft" und damit von „schlimme[n] oder scheinbar wichtige[n]" unterschieden habe. S. Freud verweist auf Tendenzen, verschieden lange Namen unterschiedlich zu behandeln: „Bekanntlich neigt man gerade bei einsilbigen Familiennamen besonders dazu, den Vornamen mitzunennen" (Freud, 1992:37).

Wie lang dürfen nun Wörter überhaupt sein? Beschränkt man den Blick einmal auf das Deutsche, so kann man sagen, daß es dafür natürlich keine theoretisch begründbare Obergrenze gibt, da die Grammatik immer zuläßt, zu einem vorhandenen Wort durch Derivation oder Komposition eines zu bilden, das um ein Element länger ist. Praktisch sieht es jedoch anders aus: Schaut man etwa in die Untersuchungen dieses Bandes, so findet man je nach Sprachstil relativ oft noch sechs- bis achtsilbige Wörter; längere kommen vor, aber nur vereinzelt; nur ganz selten trifft man auf wirklich lange Wörter wie den immer wieder erwähnten

„Donaudampfschiffahrtskapitänsanwärter", dem man ohne weiteres ein „-mütze" und mehr anhängen kann. Boettcher u.a. (1983:130) zitieren Karl Valentin mit „Isopropilprophemilbarbitursäurephenildimethildimenthylaminophirazolon", womit der Kabarettist sich über Arzneimittelbezeichnungen lustig gemacht habe. Nicht ganz so lang ist ein Ausdruck, den die australische Fluglinie Quantas in einer Werbeanzeige der letzten Jahre als Überschrift benutzt hat: „Der Südostasienaustraliensicherheitsserviceexperte". Offensichtlich wird also schon die Länge an sich als aufmerksamkeitsheischend und damit als werbewirksam verstanden.

Wörter gewinnen mit wachsender Länge aber nicht nur an Auffälligkeit; sie bereiten andererseits dem Sprecher oder Hörer (Schreiber oder Leser) größere Verarbeitungsprobleme als kürzere. Dieser Effekt wird seit langem in der Lesbarkeitsforschung ausgenutzt, wie sich in den entsprechenden Formeln, etwa dem bekannten Reading Ease-Index (RE) Fleschs ausdrückt: RE = 206.835 - 0.846 wl - 1.015sl (wl: Wortlänge; sl: Satzlänge), der für englische Texte entwickelt wurde. (Zur Forschung dazu vgl. Groeben, 1982:176ff) Daß die Wortlänge tatsächlich ein Verarbeitungsproblem darstellt, bestätigt auch Leuninger (1989:102), die als Leitsymptom der Leitungsaphasie „eine extrem starke Beeinträchtigung des Nachsprechens [bestimmt], wobei die Schwierigkeit proportional zur Länge der Wörter und Sätze zunimmt."

Eine große Rolle kommt der Wortlänge auch in der (linguistischen ebenso wie der literarischen) Stilistik zu. Das oben zitierte Beispiel Karl Valentins deutet an, daß die schiere Länge - neben andern Aspekten - ein auffälliges Merkmal von Fachsprachen sein kann. Ein nicht ganz so langes, dafür aber „echtes" Beispiel nennt Gross (1988:187): „Hochleistungsultrakurzwellengeradeausempfänger". Gelegentlich sprengen die Benennungsbedürfnisse in den Fachsprachen die Wortgrenze und es kommt zu „Mehrworttermini", „die ein gewisses Anfangsstadium der Terminusbildung charakterisieren ... und später eine Kürzung [erfahren], weil ihre Länge und Kompliziertheit im Widerspruch zur Sprachökonomie steht" (Hoffmann, ^{2}1985:171).

Auf stilistische Effekte der Wortlänge in der Literatur weist Gumppenberg (^{15}1971:63) mit seiner Parodie „Sommermädchenküssetauschelächelbeichte" „nach O. J. Bierbaum und anderen Wortkopplern" hin, die voll von entsprechend konstruierten überlangen ad hoc-Bildungen ist. Daß und wie man literarischen Stil auch mathematisch behandeln kann, zeigt W. Fucks (1968), wenn er Fragen wie die Identifikation anonymer Verfasser oder die stilistische Charakterisierung verschiedener Autoren behandelt.

Nun kann man natürlich fragen, wozu dies gut sein soll. W. Fucks plädiert für seinen Ansatz mit dem Argument: „Wer immer irgend etwas richtig zählt oder mißt, ... gewinnt allemal objektive Erkenntnisse. 'Objektiv' soll hier heißen: mitteilbar mit Zustimmungszwang" (Fucks, 1968:8). „Glückliche Fragen und fruchtbare Hypothesen" vorausgesetzt, „verknüpfen sich die Ergebnisse von Beobachtung und Experiment zu einem Denkmodell, zu einer Theorie", die ihrerseits Prognosen ermöglicht. „Man kann das nie Erfahrene wissen und das noch nie Gewesene machen" (Fucks, 1968:9).

Mit diesem Konzept erweist sich W. Fucks als einer der Wegbereiter der quantitativen Linguistik, obwohl er sich keineswegs auf linguistische Themen beschränkt. Dabei sind „Gesetz", „Theorie", „Erklärung" und „Prognose" die Schlüsselbegriffe, die andeuten, um was es eigentlich geht.

Welche Rolle spielt in diesem Zusammenhang nun die Wortlänge? Mindestens in folgenden Forschungsbereichen ist sie von Bedeutung:

a) Bereits Zipf hat entdeckt, daß der Häufigkeitsrang (r) eines Wortes in einer Frequenzliste multipliziert mit seiner Frequenz (f) annähernd eine Konstante (C) ergibt ($r \cdot f = C$), und ist damit in linguistische Handbücher vorgedrungen (Crystal, 1993:87). Diese Erkenntnis hat der Linguistik einen der ersten Kandidaten für ein „echtes" Sprachgesetz eingetragen.

b) In der Sprachtypologie wird seit Greenbergs bahnbrechender Arbeit (1954/1960) mit einer Reihe von Indizes für die Charakterisierung von Sprachen gearbeitet, von denen der sog. Synthese-Index ($S = M/W$; S: Synthesegrad, M: Zahl der Morpheme; W: Zahl der Wörter) direkt die Wortlänge mißt. Daß zwischen diesem und den meisten andern Indizes Interaktionen bestehen, haben Altmann & Lehfeldt (1973:44) mit ihrem „Graph der Merkmalszusammenhänge" demonstriert, der zeigt, wie die Indizes durch positive oder negative Korrelationen miteinander verbunden sind. Hier deutet sich bereits so etwas wie ein Konstruktionsprinzip der Sprachen an.

c) Eine weitere Form dieses Konstruktionsprinzips wurde von Köhler (1986:74) im Rahmen seiner Untersuchungen zur deutschen Lexik entwickelt. Neu an Köhlers Konzept ist der Versuch, das Zusammenspiel zwischen den Bedürfnissen der Sprachbenutzer („Systembedürfnisse") und verschiedenen Eigenschaften der Sprachen („Systemgrößen") funktional zu bestimmen. Man sieht in seinem Modell, daß die Wortlänge abhängig ist von der Phonemzahl der betreffenden Sprache, der Größe ihres Lexikons und der Frequenz der Lexeme, aber eben auch von den Bedürfnissen der Sprachgemeinschaft wie etwa dem Bedürfnis des Sprechers oder Schreibers, den Kodierungsaufwand bei der Sprachproduktion zu verringern, dem das Bedürfnis des Hörers oder Lesers widerspricht, die Dekodierungsarbeit möglichst gering zu halten, so daß beide Bedürfnisse sich aufeinander einpegeln müssen. Die Wortlänge ihrerseits aber wirkt sich auf die Mehrdeutigkeit der Wörter aus: Je länger die Wörter sind, desto spezifischer ist ihre Bedeutung, und umgekehrt. Insgesamt ergibt sich ein Modell, in dem jede Einheit einen bestimmten Platz einnimmt und an der Gestaltung des gesamten Regelkreises als zugleich beeinflußte und beeinflussende Größe beteiligt ist. Hierin ist auch ein Grund dafür zu sehen, warum seit Flesch immer wieder Wort- und Satzlänge als Kriterium für die Schwierigkeit von Texten verwendet werden:

Wenn man diese beiden Größen mißt, erhebt man mittelbar alle die anderen Größen mit, die mit ihnen in Wechselbeziehung stehen.

d) Als einen speziellen Aspekt eines solchen Regelkreises kann man das Menzerath-Altmannsche Gesetz auffassen, das einen Zusammenhang zwischen der Größe einer Einheit und der Größe ihrer Bestandteile konstatiert: Je größer das Ganze, desto kleiner seine unmittelbaren Konstituenten. Auf Wörter angewandt: Je größer die Wörter sind, desto kleiner sind ihre Silben oder Morphe, wobei die Größe der Wörter danach gemessen wird, aus wievielen Silben oder Morphen sie bestehen. Das Menzerath-Altmannsche Gesetz hat eine Vielzahl von Überprüfungen bestanden (Altmann & Schwibbe, 1989); für deutsche Wörter hat Gerlach (1982) seine Gültigkeit bestätigt; es gilt auch auf Satzebene (Köhler, 1982; Heups, 1983). Einen Menzerathschen Zusammenhang zwischen Wort und Silbe bestätigen Fenk-Oczlon & Fenk (1995:232).

e) Der Zusammenhang von Wortlänge und Frequenz hat nun schon mehrmals eine Rolle gespielt, und zwar beim Zipfschen Gesetz und in Köhlers Regelkreis. Fenk-Oczlon (1991:390) gibt auch eine Begründung dafür: „Hohe Frequenz ist deshalb ein ausgezeichneter Prädiktor für kürzere Kodierung, Priorität in Binomialen, unregelmäßige Kodierung ... etc., weil sie auch ein ausgezeichneter Prädiktor für geringe kognitive Kosten ist." [„Binomiale" (oder: „Freezes") sind feste koordinierte Syntagmen des Typs „Lust und Laune", und die entsprechende Hypothese besagt, daß in solchen Phrasen das häufigere Element meist die erste Position einnimmt (Fenk-Oczlon, 1989, 1991:384)].

f) In den Arbeiten dieses Sammelbandes geht es um einen weiteren Aspekt der Wortlänge, und zwar um die Frage, mit welcher Häufigkeit Wörter verschiedener Länge in abgeschlossenen, vollständigen Texten vorkommen und ob sich dafür Modelle entwickeln lassen. Diesen Modellen soll aber nicht nur eine rein deskriptive Qualität zukommen; sie sollen vielmehr Gesetzesstatus erreichen, um einem wissenschaftstheoretischen Prinzip Genüge zu leisten, das lautet: „Everything abides by laws" (Bunge, 1977:17). Als Gesetze müssen sie bestimmten Anforderungen genügen: Sie müssen u.a. 1. theoretisch begründbar sein und sich 2. bei empirischen Überprüfungen bewähren.

Um diesen Anforderungen zu entsprechen, kann man sich auf die von Grotjahn & Altmann (1993), Wimmer u.a. (1994) sowie von Wimmer & Altmann (1996) entwickelte Theorie stützen, die ihrerseits an Fucks (1955) und Grotjahn (1982) anknüpfen. Die Autoren gehen von der Annahme aus, daß in einem Text die Häufigkeit der Wörter der Wortlänge P_{x+1} abhängig ist von der Häufigkeit der Wörter der Wortlänge P_x; also etwa: die Zahl der zweisilbigen Wörter eines Textes richtet sich nach der Zahl der einsilbigen Wörter; gibt es viele einsilbige Wörter, wird der Anteil der zweisilbigen geringer sein als in einem andern Text, der bei gleicher Gesamtwortzahl weniger einsilbige Wörter enthält. Dies gilt analog auch für die andern Wortlängenklassen innerhalb eines Textes. Da die Relation zwischen den Wortlängenklassen nicht als konstant anzusehen ist, kann

man ansetzen:

$$P_x = g(x)P_{x-1}$$

Die Funktion $g(x)$ kann nun unterschiedliche Formen annehmen, wie inzwischen aus der Untersuchung von 35 Sprachen (Best & Altmann, 1996) bekannt ist. Setzt man $g(x) = a/x$, so erhält man die Poisson-Verteilung, die bereits Fucks (1955) vorgeschlagen hatte; mit $g(x) = (a + bx)/cx$ kommt man zu Grotjahns Modell der negativen Binomialverteilung und mit $g(x) = a/(c + x)$ zur Hyperpoisson-Verteilung (Wimmer u.a., 1994:102). Damit sind drei der wichtigsten Verteilungen benannt, die sich immer wieder bei Wortlängen bewährt haben. Die Perspektive hat sich allerdings im Lauf der Arbeiten etwas geändert. Nach Grotjahns Vorgaben konnte angenommen werden, daß die negative Binomialverteilung wenn nicht das einzige, so doch das bevorzugte Verteilungsmodell sein sollte. Inzwischen scheint die Hyperpoisson-Verteilung (Best & Altmann, 1996:88) diese Rolle zu übernehmen, und zwar aus 2 Gründen: 1. Sie ist anscheinend das beste Modell für alte Sprachen, wie die Arbeiten zu Altgriechisch, Althebräisch und Latein in diesem Band sowie zu Altisländisch (Best, 1996) zeigen. 2. Die Hyperpoisson-Verteilung bewährt sich auch bei einer großen Zahl von Sprachen der Gegenwart, wie ein Blick in weitere Arbeiten des Projekts zeigt; kein Verteilungsmodell hat sich auch nur annähernd auf so viele verschiedene Texte in unterschiedlichen Sprachen anwenden lassen. Einschränkend ist zu sagen, daß bei althochdeutschen (Best, 1996a) und mittelhochdeutschen Texten (Best, in diesem Band) die Poisson-Verteilung etwas bessere Ergebnisse ermöglicht als die Hyperpoisson-Verteilung.

g) Eine der Arbeiten dieses Bandes (Niehaus) befaßt sich mit der Modellierung von Satzlängenverteilungen. Mangels einer speziellen Theorie wurde angenommen, daß Satzlängen sich innerhalb von abgeschlossenen Texten prinzipiell genauso verhalten wie Wortlängen. Die Untersuchung von Niehaus zeigt, daß zumindest vorläufig nichts dagegen spricht, daß diese Hypothese berechtigt ist.

h) Nachdem Wort- und Satzlänge sich offenbar gleich verhalten, lag es nahe, auch weitere Spracheinheiten zu bearbeiten. Erste Versuche mit Morph- und Silbenlängen in 21 kurzen Pressetexten (Meldungen) erbrachten entsprechende Ergebnisse (Best, 1997); für die Morphlängen im Deutschen erwies sich wieder die Hyperpoisson-Verteilung als das optimale Modell; Silben folgen dagegen der Conway-Maxwell-Poisson-Verteilung, die bisher bei den andern Spracheinheiten im Deutschen keine Rolle gespielt hat. Es könnte sich hier andeuten, daß die „Silbe" sich als Sprecheinheit etwas anders verhält als die zeichenhaften Einheiten „Morph", „Wort" und „Satz". Aber auch die Conway-Maxwell-Poisson-Verteilung gehört zu der kleinen Gruppe von Verteilungen mit einer Poisson-Komponente, die durch die Theorie begründet sind.

Überblickt man die bisherigen Ergebnisse, so kann festgestellt werden, daß an

fast alle Texte und Textsorten in den 35 bisher bearbeiteten Sprachen Modelle aus dem Kreis der theoretisch begründeten Verteilungen angepaßt werden können. Lediglich zwei Textsorten wurden gefunden, die vorerst noch Probleme bereiten: chinesische Texte mit hohem fachsprachlichen Anteil (ein Beipiel dazu findet sich in Best & Zhu, 1994:28) und lappische Pressetexte (Bartens & Best, 1997); in beiden Sprachen entsprechen andere Textsorten jedoch der Hypothese. Für die weitere Arbeit lassen sich folgende Perspektiven benennen:

1. Im Deutschen soll wenigstens für einzelne Zeitabschnitte eine breitere Textsortenstreuung als die bisher erreichte angestrebt werden, damit der Frage nachgegangen werden kann, ob für verschiedene Textsorten auch unterschiedliche Modelle verwendet werden müssen bzw. in welchem Maße dies erforderlich sein wird. Außerdem muß auch dem möglichen Einfluß des Autorenstils mehr Beachtung geschenkt werden.
2. Wenn möglich, sollen weitere Sprachen bearbeitet werden, da natürlich 35 Sprachen alles andere als repräsentativ für die Sprachen insgesamt sind.
3. Da für viele Sprachen nur wenige Texte und Textsorten untersucht wurden, ist auch innerhalb der schon einbezogenen Sprachen eine weitere Streuung nach Textsorten wünschenswert.
4. Neben den Wörtern soll den andern Sprachgrößen in Zukunft mehr Aufmerksamkeit gewidmet werden, um herauszufinden, ob das Gesetz der Wortlängenverteilung sich tatsächlich, wie vermutet, als ein Gesetz der Verteilung der Längenklassen sprachlicher Einheiten überhaupt erweisen wird.
5. Als weiteres Thema soll der Frage nachgegangen werden, ob das Zusammenspiel zwischen Rang und Frequenz von Wörtern in Texten dem Zipf-Mandelbrot-Gesetz folgt, wie dies Altmann (1988:69ff) begründet und Uhlířová (1996) um einen weiteren Aspekt bereichert hat, indem sie vorschlägt, daß das Zipf-Mandelbrot-Gesetz nicht nur für ganze Texte, sondern auch für die Verteilung der Wörter innerhalb der einzelnen Wortlängenklassen eines Textes gelten sollte.

Es ist zu hoffen, daß sich möglichst viel von diesem Programm realisieren lassen wird. In allen genannten Fällen geht es im Grunde immer um die bereits zitierte These Bunges: „Everything abides by laws." Bisher gibt es keinen Grund, daran zu zweifeln.

Zum Abschluß bleibt nur noch, allen Beteiligten für ihre Mitarbeit an diesem Sammelband und die erwiesene Geduld zu danken.

Weitere Informationen zum Projekt findet man im Internet unter der Adresse: http://www.gwdg.de/~kbest/projekt.htm.

Duderstadt, im Mai 1997

Literatur

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G., & Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Altmann, G., & Schwibbe, M. H. (Hg.) (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim, Zürich, New York: Olms.

Bartens, H.-H., & Best, K.-H. (1997). Word Length Distribution in Sami Texts. In G. Altmann, J. Mikk, P. Saukkonen & G. Wimmer (Hg.), *Linguistic Structures. to Honor Juhan Tuldava*. Lisse. (erscheint).

Best, K.-H. (1996). Word Length in Old Icelandic Songs and Prose Texts. *Journal of Quantitative Linguistics*, 3, 97-105.

Best, K.-H. (1996a). Zur Bedeutung von Wortlängen, am Beispiel althochdeutscher Texte. *Papiere zur Linguistik*, 55, 141-152.

Best, K.-H. (1997). *Untersuchungen zur Verteilung von Morph- und Silbenlängen in Pressemeldungen*. In Arbeit.

Best, K.-H., & Altmann, G. (1996). Project Report. *Journal of Quantitative Linguistics*, 3, 85-88.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio linguae II* (S. 19-30), Stuttgart: Steiner.

Boettcher, W., Herrlitz, W., Nündel, E., & Swittalla, B. (1983). *Sprache*. Braunschweig: Westermann.

Bunge, M. (1977). *Treatise on Basic Philosophy. Vol.3. Ontology: The Furniture of the World*. Dordrecht: Reidel.

Crystal, D. (1993). *Die Cambridge Enzyklopädie der Sprache*. Frankfurt, New York: Campus.

Fenk-Oczlon, G. (1989). Word Frequency and Word Order in Freezes. *Linguistics*, 27, 517-556.

Fenk-Oczlon, G. (1991). Frequenz und Kognition - Frequenz und Markiertheit. *Folia Linguistica*, XXV, 361-394.

Fenk-Oczlon, G., & Fenk, A. (1995). Selbstorganisation und natürliche Typologie. *Sprachtypologie und Universalienforschung*, 48, 223-238.

Freud, S. (1904/1992). *Zur Psychopathologie des Alltagslebens*. Frankfurt: Fischer.

Fucks, W. (1955). Theorie der Wortbildung. *Mathematisch-physikalische Semesterberichte*, 4, 195-212.

Fucks, W. (1968). *Nach allen Regeln der Kunst*. Stuttgart: Deutsche Verlagsanstalt.

Gerlach, R. (1982). Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In W. Lehfeldt & U. Strauss (Hg.), *Glottometrika 4* (S. 95-102), Bochum: Brockmeyer.

Greenberg, J.H. (1954/1960). A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics*, 26, 178-194.

Groeben, N. (1982). *Leserspychologie: Textverständnis - Textverständlichkeit*. Münster: Aschendorff.

Gross, H. (1988). *Einführung in die germanistische Linguistik*. München: iudicium.

Grotjahn, R. (1982). Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift fhr Sprachwissenschaft*, 1, 44-75.

Grotjahn, R., & Altmann, G. (1993). Modelling the Distribution of Word Length: Some Methodological Problems. In R. Köhler & B. Rieger (Hg.), *Contributions to Quantitative Linguistics* (S. 141-153), Dordrecht, Boston, London: Kluwer.

Gumppenberg, H. von ([15]1971). *Das Teutsche Dichterroß*. München: dtv.

Heups, G. (1983). Untersuchungen zum Verhältnis von Satzlänge und Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In R. Köhler & J. Boy (Hg.), *Glottometrika 5* (S. 113-133), Bochum: Brockmeyer.

Hoffmann, L. ([2]1985). *Kommunikationsmittel Fachsprache*. Tübingen: Narr.

Köhler, R. (1982). Das Menzerathsche Gesetz auf der Satzebene. In W. Lehfeldt & U. Strauss (Hg.), *Glottometrika 4* (S. 103-113), Bochum: Brockmeyer.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Leuninger, H. (1989). *Neurolinguistik*. Opladen: Westdeutscher Verlag.

Paul, J. ([2]813/1996). Vorschule der Asthetik. In J. Paul. *Sämtliche Werke*. Abt. I, Bd. 5. Frankfurt: Zweitausendeins (Nachdruck der Hanser-Ausgabe).

Uhlířová, L. (1996). On the Generality of Statistical Laws and Individuality of Texts. A Case of Syllables, Word Forms, their Length and Frequencies. *Journal of Quantitative Linguistics*, 2, 238-247.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Glottometrika 16, 1997, 1-15

Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten

Karl-Heinz Best

0. Befaßt man sich mit der Frage nach den Häufigkeiten, mit denen Wörter unterschiedlicher Länge in Texten oder im Lexikon einer Sprache auftreten, so empfindet mancher eine derartige Aufgabenstellung als unüberbietbar trivial. Und in der Tat: Welche erfaßbare Eigenschaft von sprachlichen Einheiten mutet uns auf den ersten Blick als noch oberflächlicher, zufälliger und damit nichtssagender an als gerade ihre Länge und deren Häufigkeit?

Dieser Eindruck täuscht natürlich, wie sich leicht zeigen läßt. Bei genauerem Hinsehen erweist sich die Wortlänge und ihre Häufigkeit nämlich als eine Größe von hoher Relevanz für die Sprachstruktur, -verwendung und darüber hinaus u.a. für die Psycholinguistik, die Namenskunde etc.

So ist nachweisbar, daß die Länge der Wörter in einem gesetzmäßigen Zusammenhang mit ihrer jeweiligen Häufigkeit steht (Zipf, 1935:26-28; vgl. auch König, 1982:104ff). Bürmann, Frank & Lorenz (1963:74) formulieren dazu ihr Prinzip 5: „Der Erwartungswert der Wortlänge (d.h. die durchschnittliche, durch die relativen Worthäufigkeiten gewichtete Wortlänge) ist möglichst klein." Gerlach (1982) hat nachgewiesen, daß die Länge der Wörter in einem gesetzmäßigen Zusammenhang mit der Größe ihrer Konstituenten steht; Altmann, Beőthy & Best (1982), Sambor (1984) & Fickermann, Markner-Jäger & Rothe (1984) haben den Zusammenhang zwischen der Wortlänge und der Anzahl der im Lexikon verzeichneten Bedeutungen der betreffenden Wörter für etliche Sprachen belegen können. Köhler (1986:74) und Hammerl (1991:219) endlich haben ganze Regelkreise entwickelt, in denen die Wortlänge als eine Größe sich in Interaktion mit einem Komplex anderer Faktoren und Größen befindet.

Auf die naheliegende Rolle der Wortlänge für Zwecke der Sprachtypologie, der linguistischen Statistik und der Verständlichkeitsforschung wurde bereits hingewiesen (Best & Zhu, 1994:19). Im Vergleich von elf Sprachen (Fucks, 1968:80 91) zeigt sich, daß die Wortlänge keineswegs ein chaotisches Sprachmerkmal ist (vgl. hierzu auch Meinhold & Stock, 1982:204ff).

Bei seinem Versuch, Textgenres mit Hilfe einer Art von Profilvektor zu modellieren, hat Mistrík (1973:121, passim) sich neben etlichen anderen Kriterien auch auf die Wortlänge gestützt.

Für die aufschlußreiche Rolle von Wortlängen gibt es weitere Hinweise, von denen noch einige erwähnt seien. Folgt man etwa Werner (1989:40f), so ist die Suppletion typischerweise mit besonderer Kürze der beteiligten Wörter verbunden. Zur Verwendung von Vornamen im Deutschen teilt Seibicke (1982:104ff) mit, daß es offenbar eine Tendenz zu etwas längeren Ausdrücken für weibliche als für männliche Personen gibt. Die allgemeine Tendenz des Deutschen, in Texten am häufigsten einsilbige Wörter einzusetzen, trifft allerdings für die Vornamen beider Geschlechter nicht zu; es dominieren vielmehr zwei- und dreisilbige Formen, während einsilbige Ausdrücke deutlich weniger gewählt werden. Im Sprachsystem wiederum zeigt sich eine Bevorzugung zweisilbiger Lexeme (Altmann, Beöthy & Best, 1982:539).

In der Psycholinguistik gibt es Hinweise darauf, daß die „Wortlängen als lautsymbolisches Darstellungsmittel" (Ertel, 1969:133) dienen können: Wörter, die in einem semantischen Differential einen positiven Faktor „Erregung" aufweisen (E^+-Wörter), sind statistisch signifikant häufiger die längeren Items von Wortpaaren, deren Gegenstücke diesen Faktor nicht aufweisen.

In südostasiatischen, polynesischen Sprachen und im Japanischen schließlich sind die höflichen pronominalen Appellative (du, Sie, ...) länger als die weniger höflichen.

Diese Beispiele sollten genügen, um anzuzeigen, daß die Wortlänge und ihre Häufigkeit alles andere als ein lediglich triviales Merkmal von Sprachen und Texten darstellt, daß sie vielmehr als Schlüsselgrößen für Sprachstruktur und -verwendung aufgefaßt werden muß.

1. In der vorliegenden Untersuchung geht es nun darum, der Frage nachzugehen, wie häufig Wörter unterschiedlicher Länge in deutschsprachigen Pressetexten verwendet werden. Damit wird nach den Forschungen von Fucks (1968:32-38), Grotjahn (1982) und Best & Zhu (1994) zu literarischen Texten, Brief- und Sachbuchtexten sowie Frequenzwörterbüchern mit der Pressesprache ein weiterer Funktionalstil in die Untersuchungen einbezogen. Wie schon in Best & Zhu (1994:21) soll überprüft werden, ob Pressetexte der 0-gestutzten (positiven) negativen Binomialverteilung folgen, die wie folgt lautet:

$$P_x = \frac{\binom{k+x+1}{x} p^k q^x}{1-p^k} \quad , x = 1, 2, \ldots; \; k > 0; \; 0 < p < 1; \; q = 1-p.$$

Diese Verteilung hat sich zur Modellierung von Kurzprosatexten bewährt (Best & Zhu, 1994:21ff; zur theoretischen Begründung vgl. Altmann & Best, 1996); es ist aber damit zu rechnen, daß verschiedene Textsorten bzw. Funktionalstile unterschiedlichen Modellen folgen, wie Untersuchungen zum Slowakischen (Nemcová & Altmann, 1994) und Italienischen (Gaeta, 1994) gezeigt haben. Die Prüfkriterien stimmen mit denen in Best & Zhu (1994:21) überein. Zu jeder Tabelle werden die Parameter k, p, sowie das Chiquadrat (X^2) angegeben. P ist die Überschreitungswahrscheinlichkeit des entsprechenden Chiquadrats; der Index des Chiquadrats gibt die Freiheitsgrade an. Ist $P > 0.05$, gilt die Anpassung als gut. In den beiden Fällen, in denen $P < 0.05$, wurde als Maß der Abweichung der Koeffizient $C = X^2/N$ berechnet; die Anpassung gilt als gut, wenn $C \leq 0.02$ (zur Problematik des X^2-Tests vgl. Best & Zhu, 1994:21 bzw. Grotjahn & Altmann, 1993).

2. Die Datenaufnahme entspricht der in Best & Zhu (1994) verwendeten. Da bei Zeitungstexten aber einige Probleme mehr auftreten als bei Kurzprosatexten, sollen hier die Prinzipien der Datenaufnahme genauer beschrieben werden.

Zunächst: Es wurde immer nur der laufende Text mit Überschrift, aber ohne alle sonstigen Zusätze wie Autorenname, Bildunterschriften etc. bearbeitet. Inzwischen gibt es Überlegungen, daß eine Beschränkung auf den laufenden Text allein günstiger sein könnte; die hier vorgenommene Art der Datenerfassung hat sich aber im vorliegenden Fall nicht als störend erwiesen und wurde deshalb nicht mehr nachträglich korrigiert.

Als „Wort" wurde wieder das orthographische Wort gewählt; die Zahl der Silben wurde nach der Zahl der Vokale bzw. Diphthonge im Wort bestimmt. Problematische Formen wie Abkürzungen, Akronyme etc. wurden in der Form ausgewertet, die man benutzt, wenn man die betreffenden Wörter laut liest. Also z.B. „FDP": 1 Wort, 3 Silben ([efde:pe:]).

Um eine möglichst einfache Operationalisierung bei der Bestimmung der Silbenzahl zu gewährleisten, wurden Fälle wie „Studie" ([stu:diə]), „Union" ([unio:n]) und „sozial" ([zotsia:l]) buchstabengetreu als dreisilbig gewertet. Dies entspricht der Bühnenaussprache, bei der unsilbische Vokale in der Regel als silbentragende realisiert werden (Duden: Aussprachewörterbuch, 53).

Zahlwörter wurden entsprechend den orthographischen Konventionen im Deutschen als ein Wort gewertet. So wird „1993" als „neunzehnhundertdreiundneunzig" realisiert, wenn es sich um eine Jahreszahl handelt, oder als „(ein)tausendneunhundertdreiundneunzig", wenn es um eine Entfernungsangabe, einen Geldbetrag oder dgl. geht. Diese Art der Wertung hat sich inzwischen bei weiteren Untersuchungen als problematisch erwiesen, wenn solche Zahlwörter in Texten gehäuft vorkommen. In der vorliegenden Untersuchung wurde aber keine nachträgliche Korrektur vorgenommen, da Zahlwörter in den untersuchten Texten nur vereinzelt vorkommen und daher außer in einem der Texte (s.u.) fast kei-

ne Rolle spielen. Zahlen wie „19,93“ (Liter, Kilometer etc.) wurden entsprechend der gelesenen Form („neunzehn Komma/Punkt dreiundneunzig“) als aus drei Wörtern bestehend gewertet. Wieder anders muß man Telefonnummern („1993“ = „neunzehn dreiundneunzig“ oder „eins neun neun drei“: 2 bzw. 4 Wörter) bearbeiten.

Durch Kürzung von „es“ zu „'s“ entstehen im Deutschen bisweilen nullsilbige Wörter; sie spielen in den Pressetexten dieser Arbeit aber fast keine Rolle. Formen wie „ins“, „ans“ etc. sind dagegen bereits orthographisch als ein Wort ausgezeichnet. Die beiden einzigen Vorkommen von „'s“ wurden entsprechend als Bestandteile ihrer Nachbarwörter betrachtet und nicht eigens als nullsilbige Wörter ausgewiesen.

Der Bindestrich wird ebenso wie der Trennungsstrich als Hinweis auf die Einheit eines Wortes aufgefaßt. Lediglich "free compounding" (Rothe, 1988: 122; vgl. Pilz, 1981:93: „Juxtaposita“), das im Deutschen vorwiegend bei Namenkomposita („Dugena Fachgeschäfte“, „Waldbaur Schokolade“) vorkommt, sich aber allmählich auch darüber hinaus ausbreitet („das Original Teil“: VW-Werbung, Stern 43/1989, 115), wird wegen des Kriteriums „orthographisches Wort“ als mehrwortiges Syntagma behandelt.

Es ist klar, daß eine schematische Übertragung dieser Entscheidungen auf andere Sprachen sich von selbst verbietet: Es wäre wenig sinnvoll, im Französischen „voulez-vous“ nur wegen des Bindestrichs als ein Wort oder im Englischen "Jack's" wegen des Apostrophs als zwei Wörter zu werten. Die hier vorgestellten Entscheidungen müssen also bei entsprechenden Untersuchungen in anderen Sprachen an die Konventionen der betreffenden Sprache angepaßt werden.

3. Die Textauswahl ist relativ willkürlich getroffen worden. Um eine einigermaßen homogene Textgruppe zu erhalten, wurden nur Organe der Tages- und Wochenpresse berücksichtigt. Diese wenden sich an ein breites Leserpublikum ohne sich auf allzu spezielle Interessen einstellen zu müssen. Die fachsprachlichen Einflüsse, die sich ja auch in der Wortlänge niederschlagen, sollten in solchen Presseorganen also geringer sein, als in denen der Hobby-, Fach-, und Wissenschaftspresse.

Für diese Untersuchung wurden Texte des österreichischen Wochenmagazins „profil“ sowie des „Eichsfelder Tageblatts“ (ET), einer Regionalausgabe des Göttinger Tageblatts mit eigener Lokalredaktion, ausgewählt. (Daten zu 10 Texten aus „Der Spiegel“ finden sich in Altmann & Best, 1996.) Wenn hier speziell auch ausgesprochen kurze Texte bearbeitet wurden, so deshalb, um Aufschluß darüber zu gewinnen, wo etwa eine untere Grenze für die Länge von Texten anzusiedeln ist, die für Wortlängenzählungen ausgewertet werden können. Es könnte ja immerhin sein, daß Texte eine gewisse Mindestlänge haben müssen, damit sich eine gesetzmäßige Wortlängenverteilung einstellen kann. Auch nach

dieser Untersuchung läßt sich aber keine untere Grenze für solche Analysen angeben.

4. Die Datenerhebung hat folgende Ergebnisse erbracht:

Pressetexte aus Österreich

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	181	178.76	322	332.24	282	269.54
2	125	126.48	266	239.40	171	190.13
3	70	74.80	108	122.97	101	108.91
4	46	39.89	50	50.40	81	57.58
5	22	19.88	14	17.56	29	29.16
6	5	9.44	7	5.38	10	14.38
7	5	4.33	1	1.49	5	6.97
8	3	3.42	1	0.38	3	3.33
9	-	-	1	0.14	0	1.57
10	-	-	-	-	0	0.74
11	-	-	-	-	0	0.34
12	-	-	-	-	0	0.16
13	-	-	-	-	0	0.07
14	-	-	-	-	0	0.03
15	-	-	-	-	1	0.09
	$k = 2.9418$; $p = 0.6010$; $X^2_5 = 3.76$; $P = 0.58$.		$k = 13.4381$; $p = 0.9002$; $X^2_4 = 6.81$; $P = 0.15$.		$k = 1.6021$; $p = 0.5597$; $X^2_7 = 16.22$; $P = 0.02$; $C = 0.02$.	

Text 1: H. Lackner, Gesprächige Trappisten (profil 24. Jg., Nr. 3/1993:21; Sparte „Österreich“).
Text 2: E. Menasse, Die flüchtige Zeugin (profil 24. Jg., Nr. 4/1993:31-32; Sparte „Österreich“).
Text 3: G. Mayr, O. Tanzer: Tatort Sportverein (profil 24. Jg., Nr. 6/1993:33; Sparte „Österreich“).

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	267	264.24	314	302.24	283	288.93
2	161	160.05	192	199.64	179	165.99
3	95	90.03	96	111.63	76	85.91
4	46	49.35	78	58.64	49	42.83
5	25	26.69	26	29.81	19	20.94
6	7	14.32	10	14.86	9	10.12
7	9	7.64	10	7.30	5	4.85
8	9	4.06	3	3.55	1	2.31
9	0	2.15	1	1.71	2	1.09
10	2	2.47	0	0.82	0	0.52
11	-	-	1	0.80	0	0.24
12	-	-	-	-	1	0.27
	$k = 1.1662$; $p = 0.4806$; $X^2_5 = 5.28$; $P = 0.38$.		$k = 1.4429$; $p = 0.5422$; $X^2_7 = 13.00$; $P = 0.07$.		$k = 1.2474$; $p = 0.5394$; $X^2_6 = 4.63$; $P = 0.59$.	

Text 4: Herbert. Lackner, Der Pröll-Bock (profil 24. Jg., Nr. 6/1993:24; Sparte„Österreich").
Text 5: Andreas Weber, Roter Ampelmann (profil 24. Jg., Nr. 8/1993:20; Sparte „Österreich").
Text 6: Hubertus Czernin, Herrn Haiders. Helfer (profil 24. Jg., Nr.3 993:11; „Leitartikel").

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	250	245.86	242	235.25	285	272.21
2	186	183.75	153	153.11	157	160.01
3	93	102.18	67	82.77	74	86.12
4	54	50.31	50	41.70	49	44.93
5	26	23.16	22	20.24	25	23.07
6	10	10.22	10	9.60	6	11.73
7	0	4.38	4	4.48	11	5.92
8	3	1.83	2	2.07	3	6.01
9	1	1.31	1	1.78	-	-
	$k = 2.0487$; $p = 0.6352$; $X^2_6 = 6.72$; $P = 0.35$.		$k = 1.5123$; $p = 0.5696$; $X^2_6 = 5.38$; $P = 0.50$.		$k = 1.2027$; $p = 0.5113$; $X^2_4 = 6.07$; $P = 0.19$.	

Text 7: H. Czernin, Bomben auf Belgrad (profil 24. Jg., Nr. 4/1993: „Leitartikel").
Text 8: J. Votzi, Hoffnungslicht Heide? (profil 24. Jg., Nr. 6/1993:11; „Leitartikel").
Text 9: H. Czernin, Kindesweglegung (profil 24. Jg., Nr. 8/1993:11; „Leitartikel").

	Text 10	
x	n_x	NP_x
1	313	288.24
2	149	178.99
3	88	94.88
4	62	47.42
5	28	22.98
6	6	10.92
7	4	5.13
8	2	2.38
9	1	2.06
	$k = 1.4139$; $p = 0.5608$; $X^2_6 = 16.29$; $P = 0.01$; $C = 0.02$.	

Text 10: Georg Hoffmann-Ostenhoff, Italien ist überall (profil 24. Jg., Nr. 10/1993:11; „Leitartikel").

Zur Datenaufnahme:

Es wurden keine Texte berücksichtigt, in denen mundartliche Zitate enthalten sind. Autorennamen wurden ebenso wie Cartoons, Bildunterschriften und Fußnoten nicht berücksichtigt. Bei Eigennamen mit silbischen Liquiden wie „Jandl", „Wetzl", „Mayr", „Radlbrunners" wurden die Liquidae als eigene Silbe gewertet. In „Johannes Paul II." wurde „II." entsprechend der gelesenen Form „der Zweite" als 2 Wörter mit 1 bzw. 2 Silben gewertet. „war's": 1 Wort, 1-silbig.

Lokalglossen

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	75	75.21	65	64.99	50	46.14
2	56	50.80	51	51.59	25	31.41
3	20	26.35	30	28.31	18	18.13
4	12	11.60	11	12.07	13	9.52
5	7	4.56	4	4.26	5	4.71
6	1	2.48	1	1.29]	2	2.23
7	-	-	1	0.49]	1	1.86
	$k = 5.5915$; $p = 0.7950$; $X_3^2 = 4.26$; $P = 0.24$.		$k = 25.9582$; $p = 0.9411$; $X_3^2 = 0.25$; $P = 0.97$.		$k = 2.6809$; $p = 0.6300$; $X_4^2 = 3.32$; $P = 0.51$.	

Text 1: Qual der Wahl (ET 27.4.93, Seite „Duderstadt").
Text 2: Völlig unnötig (ET 27.4.93, Seite „Duderstadt").
Text 3: Beschattet (ET 26.4.93, Seite „Duderstadt").

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	90	91.22	67	67.96	121	117.86
2	52	45.18	41	35.99	67	74.19
3	13	19.68	10	15.04	37	32.20
4	8	7.98	6	5.44	11	10.83
5	4	3.09	2	1.78]	3	3.92
6	2	1.85	1	0.79]	-	-
	$k = 2.1366$; $p = 0.6842$; $X_3^2 = 3.59$; $P = 0.31$.		$k = 4.4543$; $p = 0.8058$; $X_2^2 = 2.53$; $P = 0.28$.		$k = 28.2225$; $p = 0.9569$; $X_2^2 = 1.71$; $P = 0.43$.	

Text 4: Bitte, bitte leg nicht auf (ET 28.4.93, Seite „Duderstadt").
Text 5: „Ich nicht" (ET 29.4.93, Seite „Duderstadt").
Text 6: Heiß auf Eis (ET 30.4.93, Seite „Duderstadt").

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	78	75.87	124	121.76	72	68.51
2	37	39.84	63	68.90	39	45.74
3	20	19.88	36	34.49	24	25.25
4	10	9.66	21	16.14	21	12.48]
5	7	4.62	7	7.24	5	5.73
6	1	2.18	2	3.15	0	2.50]
7	0	1.02]	0	1.34	1	1.79]
8	1	0.93]	0	0.56]	-	-
9	-	-	0	0.23]	-	-
10	-	-	1	0.19]	-	-
	$k = 1.3486$; $p = 0.5529$; $X_4^2 = 2.58$; $P = 0.63$.		$k = 2.0574$; $p = 0.6298$; $X_4^2 = 3.24$; $P = 0.52$.		$k = 3.1608$; $p = 0.6791$; $X_1^2 = 2.14$; $P = 0.14$.	

Text 7: Kletternder Kastenklau (ET 3.5.93, Seite „Duderstadt").
Text 8: Käsepost (ET 4.5.93, Seite „Duderstadt").
Text 9: Sägen nach dem Segen (ET 6.5.93, Seite „Duderstadt").

	Text 10	
x	n_x	NP_x
1	102	102.08
2	58	52.40
3	17	25.82
4	16	12.46
5	9	5.94]
6	0	2.80]
7	2	2.50]
	$k = 1.2711$; $p = 0.5480$; $X_2^2 = 4.62$; $P = 0.10$.	

Text 10: Der Clou (ET 7.5.93, Seite „Duderstadt").

Zur Datenaufnahme:

Alle Texte ohne Verfasserkürzel und ohne Spartenüberschrift „Auf ein Wort".
ET: Eichsfelder Tageblatt.

Glosse

Text 1

x	n_x	NP_x
1	219	224.12
2	159	146.79
3	60	64.46
4	17	21.35
5	8	5.69
6	0	1.27
7	0	0.24
8	1	0.08
	$k = 170.3131$; $p = 0.9924$; $X^2_3 = 3.47$; $P = 0.32$.	

Text 1: „Große Frauen" (aus: Titanic, Satiremagazin, Nr. 4, April 1993, S.20).

Zur Datenaufnahme:

Das Gleichheitszeichen („=") wurde gemäß seiner gesprochenen Realisierung („gleich") als einsilbiges Wort gewertet.
Der Text wurde von Michael Conrad, Göttingen, bearbeitet.

Kurztexte

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	26	24.12	28	27.41	26	23.75
2	13	16.95	18	19.60	14	15.70
3	16	10.25	13	11.35	9	9.84
4	3	5.70	5	5.81	7	5.99
5	2	3.00	3	2.73	5	3.59
6	1	1.52	1	1.21	0	2.13
7	2	1.46	1	0.89	2	1.25
8	-	-	-	-	0	0.73
9	-	-	-	-	1	1.02
	$k = 2.4386$; $p = 0.5911$; $X^2_4 = 6.39$; $P = 0.18$.		$k = 3.6429$; $p = 0.6919$; $X^2_3 = 0.52$; $P = 0.91$.		$k = 1.3689$; $p = 0.4419$; $X^2_5 = 3.32$; $P = 0.65$.	

Text 1: „Noch mehr Wahlen sind ungültig" (ET 5.5.93, Titelseite).
Text 2: Chasbulatow gibt Behörden Mitschuld (ET 5.5.93, Titelseite).
Text 3: 20000 Tote in Tadschikistan (ET 5.5.93, Titelseite).

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	27	26.90	37	35.03	21	21.08
2	15	17.40	21	20.87	22	18.76
3	9	10.12	7	11.53	7	12.30
4	4	5.56	10	6.13	9	6.63
5	9	2.94	2	3.17	3	3.11
6	0	1.52	2	1.62	0	1.31
7	1	0.77	1	1.65	2	0.81
8	1	0.79	-	-	-	-
	$k = 1.8666$; $p = 0.5486$ $X^2_1 = 1.48$; $P = 0.22$.		$k = 1.5516$; $p = 0.5331$; $X^2_4 = 5.10$; $P = 0.28$.		$k = 8.4511$; $p = 0.8117$; $X^2_3 = 3.70$; $P = 0.30$.	

Text 4: Acht Jahre Haft für Allgäu-Rauschgiftboß (ET 5.5.93, „Welt im Spiegel").
Text 5: US-Weinkönig Gallo bei Unfall getötet (ET 5.5.93, „Welt im Spiegel").
Text 6: VW testet schnellen Verkehrswarnfunk (ET 5.5.93, „Welt im Spiegel").

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	43	40.94	35	33.33	20	20.38
2	22	25.72	16	18.33	17	14.13
3	14	13.42	11	9.20	4	7.38
4	8	6.28	3	4.40	4	3.22
5	3	2.74	2	2.04	1	1.24
6	1	1.90	2	1.70	1	0.65
	$k = 3.0742$; $p = 0.6915$; $X^2_3 = 1.57$; $P = 0.67$.		$k = 1.7078$; $p = 0.5937$; $X^2_3 = 1.24$; $P = 0.74$.		$k = 6.7462$; $p = 0.8209$; $X^2_2 = 2.33$; $P = 0.31$.	

Text 7: Kaum Platz für Flüchtlinge (ET 5.5.93, Titelseite).
Text 8: Wieder verschoben (ET 5.5.93, Titelseite).
Text 9: Saftig beurlaubt (ET 5.5.93, Titelseite).

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	25	27.09	14	13.89	55	55.30
2	16	12.80	12	10.13	32	30.22
3	5	5.97	3	6.33	17	17.05
4	3	2.76	5	3.63	8	9.77
5	1	1.27	2	1.96	5	5.65
6	0	0.58	1	1.02	4	3.29
7	1	0.53	0	0.51	3	1.92
8	-	-	0	0.25	2	2.80
9	-	-	1	0.28	-	-
	$k = 1.0787$; $p = 0.5454$; $X^2_3 = 1.21$; $P = 0.75$.		$k = 2.4997$; $p = 0.5832$; $X^2_4 = 2.62$; $P = 0.62$.		$k = 0.8223$; $p = 0.4001$; $X^2_5 = 1.46$; $P = 0.92$.	

Text 10: Goellner besiegt Chang (ET 5.5.93, Titelseite).
Text 11: Spion bei Thierse? (ET 7.5.93, Titelblatt).
Text 12: Süssmuth zu Europa 1993 (ET 7.5.93, Seite „Duderstadt“).

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	36	35.70	29	26.02	24	23.67
2	23	25.11	10	16.40	9	8.84
3	16	12.63	12	8.98	3	4.20
4	5	5.08	6	4.55	3	2.21
5	1	2.48	3	4.05	1	1.23
6	-	-	-	-	1	0.71
7	-	-	-	-	1	1.14
	$k = 12.7447$; $p = 0.8977$; $X^2_2 = 1.95$; $P = 0.38$.		$k = 2.3006$; $p = 0.6179$; $X^2_2 = 4.58$; $P = 0.10$.		$k = 0.0987$; $p = 0.3198$; $X^2_3 = 0.64$; $P = 0.89$.	

Text 13: Gartenlaube am Sandwasser brennt (ET 7.5.93, Seite „Duderstadt“).
Text 14: Zur LNS-Planung (ET 7.5.93, Seite „Duderstadt“).
Text 15: Boris ausgeschieden (ET 7.5.93, Titelseite).

	Text 16	
x	n_x	NP_x
1	15	14.79
2	5	6.18
3	4	3.08
4	2	1.65
5	1	0.93
6	1	1.37
	$k = 0.2680$; $p = 0.3410$; $X^2_2 = 0.66$; $P = 0.72$.	

Text 16: Es heitert auf (ET 7.5.93, Titelseite).

Zur Datenaufnahme:

Alle Kurztexte nur mit Überschrift und Text, ohne Verfasserkürzel, Ort, Verweis auf weiteren Artikel und Fotonachweis.
ET: Eichsfelder Tageblatt.

5. Als Ergebnis kann festgestellt werden: Wie schon die Kurzprosatexte (Best & Zhu, 1994) lassen sich auch sämtliche bisher untersuchten Pressetexte mit der 0-gestutzten (positiven) negativen Binomialverteilung hinsichtlich der Wortlängenhäufigkeiten modellieren. (Zwei der 36 Texte - Text 3 und 10 aus „profil" - erfüllen unsere Kriterien gerade noch oder gerade nicht mehr, je nach dem, wie man ihre Werte auf- bzw. abrundet. In Text 3 machen sich womöglich die Zahlwörter negativ bemerkbar.) Das gilt übrigens auch für die in Altmann & Best (1996) vorgestellten 26 Texte, darunter 10 weitere Pressetexte (aus „Der Spiegel"), sowie einige literarische und Brieftexte von Goethe, Schiller und Gegenwartsautoren (Daten z.T. aus Grotjahn, 1979). Auch die Kurztexte zeigen keine Abweichungen. Die 0-gestutzte (positive) negative Binomialverteilung erweist sich damit als ein gutes Modell für ein breites Textspektrum im Deutschen. Ob sie für alle Textsorten bzw. Funktionalstile des Deutschen geeignet ist, muß in weiteren Untersuchungen überprüft werden.

Noch wenig untersucht ist die Frage, ob die 0-gestutzte (positive) negative Binomialverteilung auch für ältere deutsche Texte geeignet ist. Eine erste Untersuchung (Müller, 1993) von 10 Briefen und einem weiteren Prosatext Martin Luthers ergab Wortlängenhäufigkeiten, die sich mit dieser Funktion nicht modellieren ließen.

Literatur

Altmann, G., Beőthy, E., & Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift. für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 35, 537-543.

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In P. Schmidt (Hg.), *Glottometrika 15* (S. 166-180), Trier: W.V.T.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Bürmann, G., Frank, H., & Lorenz, L. (1963). Informationstheoretische Untersuchungen über Rang und Länge deutscher Wörter. In *Grundlagenstudien aus Kybernetik und Geisteswissenschaft*, 4, 73-90.

Duden. Aussprachewörterbuch (³1990). Mannheim-Wien-Zürich: Dudenverlag.

Ertel, S. (1969). *Psychophonetik*. Göttingen: Hogrefe.

Fickermann, I., Markner-Jäger, B., & Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In J. Boy & R. Köhler (Hg.), *Glottometrika 6*, (S. 115-126), Bochum: Brockmeyer.

Fucks, W. (1968). *Nach allen Regeln der Kunst*. Stuttgart: Deutsche Verlagsanstalt.

Gaeta, L. (1994). Wortlängenverteilung in italienischen Texten. *Zeitschrift für empirische Textforschung*, 1, 44-48.

Gerlach, R. (1982). Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In W. Lehfeld & U. Strauß (Hg.), *Glottometrika 4* (S. 95-102), Bochum: Brockmeyer.

Grotjahn, R. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft*. Bochum: Brockmeyer.

Grotjahn, R. (1982). Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft*, 1, 44-75.

Grotjahn, R., & Altmann, G. (1993). Modelling the distribution of word length. In R. Köhler & B.B. Rieger (Hg.), *Contributions to quantitative linguistics* (QUALICO, Trier 1991) (S. 141-153), Dordrecht, Boston, London: Kluwer.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells*. Trier: WVT.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

König, W. (1978). *dtv-Atlas zur deutschen Sprache*. München: dtv.

Meinhold, G., & Stock, E. (²1982). *Phonologie der deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut.

Mistrík, J. (1973). *Exakte Typologie von Texten*. München: im Komm., Sagner.

Müller, B. (1993). Wortlängenhäufigkeiten in Texten Martin Luthers. *Manuskript*.

Nemcová, E., & Altmann, G. (1994). Zur Wortlänge in slowakischen Texten. *Zeitschrift für empirische Textforschung*, 1, 40-43.

Pilz, K.D. (1981). *Phraseologie*. Stuttgart: Metzler.

Rothe, U. (1988). Polylexy and compounding. In K.P. Schulz (Hg.), *Glottometrika 9* (S. 121-134), Bochum: Brockmeyer.

Sambor, J. (1984). Menzerath's law and the polysemy of words. In J. Boy & R. Köhler (Hg.), *Glottometrika 6* (S. 94-114), Bochum: Brockmeyer.

Seibicke, W. (1982). *Die Personennamen im Deutschen*. Berlin, New York: de Gruyter.

Werner, O. (1989). Sprachökonomie und Natürlichkeit im Bereich der Morphologie. *Zeitschrift. für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 42, 34-47.

Zipf, G.K. (1935). *The psycho-biology of language*. Cambridge (Mass.): The M.I.T. Press.

Glottometrika 16, 1997, 16-39

Wortlängenverteilung in deutschen Barockgedichten

Birte Christiansen

1. Gegenstand dieser Untersuchung ist die Frage, welchen Gesetzmäßigkeiten die Wortlängenhäufigkeiten in deutschen Barockgedichten folgen. Es hat sich inzwischen in einer ganzen Reihe von Sprachen und innerhalb der Sprachen in unterschiedlichen Entwicklungsphasen und Textsorten immer wieder gezeigt, daß solche Gesetzmäßigkeiten aufgedeckt werden können (vgl. etwa die Beiträge in Glottometrika 15 und 16). Dabei ist beobachtet worden, daß einerseits für verschiedene Entwicklungsstufen und Textsorten einer Sprache und andererseits für verschiedene Sprachen teils gleiche, teils verschiedene Modelle gefunden werden können. Untersuchungen zum Deutschen haben ergeben, daß ein überwiegender Teil der Texte der Gegenwartssprache der positiven (nullgestutzten) negativen Binomialverteilung folgen (vgl. Best & Altmann, 1996), während frühneuhochdeutsche Texte Luthers der 1-verschobenen Hyperpoisson-Verteilung entsprechen (Kuhr & Müller, 1997). Bei diesen Untersuchungen zum Deutschen wurden bisher leider nur wenige Gedichte berücksichtigt (Best & Altmann, 1996); diese folgen aber offenbar in der Regel dem gleichen Modell wie gleichzeitige Prosatexte verschiedenster Art.

Auf die Merkmale der Barockdichtung soll an dieser Stelle nicht näher eingegangen werden. Die Sprachgestaltung der Barockdichtung weist jedoch einige Besonderheiten auf, deren relevante Punkte in Bezug auf die Untersuchung kurz herausgehoben werden: Bei der Dichtung des 17. Jahrhunderts handelt es sich überwiegend um Gesellschaftsdichtung. Die Gelegenheitsdichtung wird erst spät durch Individualdichtung ergänzt. Die Öffentlichkeit dieser Dichtung wird entscheidend von der Rhetorik bestimmt, die verlangt, daß sich der sprachliche Ausdruck nach der zu erzielenden Wirkung zu richten habe.

Zur Zeit des Barock findet der im europäischen Vergleich verspätete Übergang von der lateinischen zur deutschen Sprache statt. Die aus dem Gelehrtenstand kommenden Dichter erneuern die volkssprachliche Dichtung auf einer humanistischen Basis. Damit ist sie für eine intellektuell-aristokratische bürgerliche Gesellschaftsschicht bestimmt.

Der Reformer Martin Opitz (1597-1639) entwickelt Muster für alle Gattungen und Formen. Die Verskunst der Lyrik wird bestimmt durch alternierende Verse und einen natürlichen Wortakzent (Betonungsgesetz).

Neben diesen formalen Vorgaben tritt eine Vielfalt an Formen und Inhalten auf (Sonett, Elegie, Ode, Epigramm; klagende und heitere Liebesdichtung, geselliges Lied, Lobdichtung, Satire).

Die Epoche kann in zwei Abschnitte geteilt werden, die sich jeweils durch bestimmte sprachliche bzw. stilistische Mittel voneinander unterscheiden:

A) *Vorbarocker Klassizismus*: Überwiegend Gesellschaftsdichtung, Einheitlichkeit in Schrift und Sprache, Überbetonung der Form und hoher Stellenwert des technischen Könnens. Vertreter: Opitz, Fleming, Gryphius.
B) *Hochbarock* (ab 1650): Zunehmend auch Individualdichtung, technisches Können „steigert" sich im Manierismus. Der Schwell- und Prunkstil dieser Phase ist gekennzeichnet durch antike Ornamente und Motive, Pathos, christlichen Gehalt und allgemeines Häufen und Anschwellen. Letzteres gelingt vor allem mit Hilfe von Worthäufungen, Wortwiederholungen, Antithetik und asyndetischen Reihen.

Die Untersuchung der Barocklyrik soll zeigen, ob sich die genannten sprachlich-stilistischen Elemente derart auf die Wortlängen auswirken, daß die Silbenauszählung deutlich verschiedene Wortlängenhäufigkeiten in den zwei Phasen der Epoche ergibt.

Für die Untersuchung von Barockgedichten eröffnen sich nun folgende Perspektiven:

1. gehören Texte des Barock von ihrem Verteilungsmodell her eher zu den frühen Phasen der Entwicklung des Deutschen oder zu den neueren? Oder bilden sie etwa eine Zwischenstufe mit eigenem Gepräge?
2. Entsprechen alle diese Gedichte, die zwar von verschiedenen Autoren stammen, aber doch immerhin alle dem Barock zuzuordnen sind, dem gleichen Verteilungsmodell, oder ist mit größeren Unterschieden zu rechnen?
3. Gibt es eine Übereinstimmung zwischen Gedichten und anderen Gattungen, oder zeigen sich selbst innerhalb eines begrenzten Zeitraums und Textsortenspektrums deutliche Unterschiede?

Als Fernziel zeichnet sich eine Charakterisierung des Deutschen sowohl in seiner historischen Veränderung als auch in seiner Variabilität in jedem beliebigen Zeitraum nach Stilen (Gattung, Textsorte, Autoren) ab; die Untersuchung von Barockgedichten ist natürlich nur ein Schritt in diese Richtung, vielleicht aber doch ein wichtiger.

2. Um Wortlängenverteilungen zu untersuchen, muß man „Wort“ und „Wortlänge“ bestimmen. Wie schon in Best & Zhu (1994:20) wird das „Wort“ als „orthographisches Wort“ aufgefaßt und die Wortlänge in der Zahl der Silben pro Wort gemessen. Dabei entspricht die Zahl der Silben im Wort der Zahl der Vokale und Diphthonge. Bei einigen Diphthongen, die eigentlich zu Monophthongen geworden sind, wurde die alte Schreibweise beibehalten. Wörter wie „Liecht“ wurden somit als einsilbig gezählt. Ähnlich sieht es bei Vokallängen aus, die durch eine Verdopplung des Vokalzeichens erzielt wurden. Auch diese Wörter, zum Beispiel „schooß“, zählen als einsilbige. Wörter mit Apostroph wie „Sonn'“, „Nacht'“ oder „Seh'“ wurden ebenfalls als einsilbig gezählt. Zu beobachten war auch das Fehlen des Vokals ‘e’ nach einem vorhergehenden Vokal in Endsilben: „Fewr“, „Trawr“, „Mawr“. Da es sich nicht um eine Abkürzung handelt und an anderer Stelle das ‘e’ geschrieben wird, zählten derartige Wörter als zweisilbig. Viele O's in einzelnen Gedichten führten dazu, daß die Zahl einsilbiger Wörter überdurchschnittlich hoch lag.

3. Um keine allzu großen sprachlichen Unterschiede zu erhalten, wurden nur Barockgedichte berücksichtigt. Die Auswahl erfolgte in diesem Rahmen mehr oder weniger willkürlich.

Die Barocklyrik ist in dem Umfang der einzelnen Gedichte durch relativ große Homogenität gekennzeichnet. Sehr viele Gedichte haben eine Länge von ungefähr 130 Wörtern. Es wurden jedoch auch einige längere Gedichte ausgezählt, wobei das längste einen Umfang von 615 Wörtern hat. Alle Gedichte haben jedoch mindestens 100 Wörter, da diese Länge als ungefähre untere Grenze für die Anzahl von Wörtern in einem Text angenommen worden ist, ohne daß es zu Homogenitätsproblemen kommt.

Es wurde versucht, möglichst eine gleiche Anzahl und einen ähnlichen Umfang von Gedichten für die beiden eingangs dargestellten Phasen des Barock zu untersuchen. Als Auswahlkriterien diente die zeitliche Einordnung der Gedichte sowie die Zuordnung der Dichter nach sprachlich-stilistischen Merkmalen ihrer Lyrik.

Mit dieser Auswahl verband sich das Interesse, die Vermutung zu belegen, daß sich die jeweils unterschiedlichen sprachlichen Elemente von vorbarockem Klassizismus und Hochbarock auf die Wortlängenhäufigkeit in den Gedichten niederschlagen.

Da für die erste Phase 45 zur Auszählung geeignete Gedichte (d.h. mit einem durchschnittlichen, vergleichbaren Umfang) von drei Autoren mit jeweils über zehn Gedichten vorlagen, jedoch für die zweite Phase die Anzahl der Gedichte pro Autor wesentlich geringer war, wurde im Interesse der Vergleichbarkeit der beiden Phasen neun Gedichte von insgesamt drei Dichtern hinzugenommen. Auch wenn die Anzahl der Gedichte der jeweiligen Autoren für den Einzelnen eventuell kein repräsentatives Bild vermitteln, tragen sie jedoch insgesamt zur

Darstellung der Gestaltungsprinzipien des Hochbarock durchaus bei. Somit wurden für diese zweite Phase insgesamt 29 Gedichte von sechs Autoren ausgezählt. Allgemein gilt jedoch, daß der Grad der Datenhomogenität wohl um so höher ist, je mehr Texte nur eines Autors ausgezählt werden.

Das Problem der Vermeidung „verschiedener Funktionalstile/Textsorten innerhalb einer Sprache“, die möglicherweise unterschiedlichen Modellen folgen, stellte sich bei der Auszählung der Barockgedichte nicht. Die ausgewählten Texte wurden mit Überschriften vollständig ausgewertet.

4. An die erarbeiteten Daten der Gedichte wurden mit dem Altmann-FITTER (1994) die positive Poissonverteilung und die nullgestutzte (positive) negative Binomialverteilung angepaßt. Es zeigte sich, daß die positive Poissonverteilung für fast alle Gedichte ein geeignetes Modell darstellt.

$$P_{xP} = \frac{e^{-a}\, a^x}{x!\,(1-e^{-a})}, \qquad x = 1, 2, 3, \ldots$$

$$P_{xNB} = \frac{\binom{k+x-1}{x} p^k q^x}{1-p^k}, \qquad x = 1, 2, 3, \ldots$$

$$P_{xVP} = \frac{e^{-a}\, a^{x-1}}{(x-1)!}, \qquad x = 1, 2, 3, \ldots$$

Grundlage für diese Bewertung sind die Prüfgrößen X^2 (Chiquadrat), P (Überschreitungswahrscheinlichkeit des Chiquadrats) und der Diskrepanzkoeffizient $C = X^2/N$. Letzterer wird dann zur Hilfe genommen, wenn P mangels Freiheitsgraden (FG) nicht bestimmt werden kann oder ein relativ hohes n (Zahl der Wörter pro Gedicht) gegeben ist. Eine Anpassung gilt als zufriedenstellend, wenn $P \geq 0.05$ oder $C \leq 0.02$. Als noch akzeptabel werden Werte von $0.01 \leq P \leq 0.05$ betrachtet.

Außer diesen Werten finden sich in den Tabellen: X: Wortlänge, n_x : Wörter der Länge x, NP_{xP}: nach der positiven Poissonverteilung berechnete Werte, NP_{xNB}: nach der positiven negativen Binomialverteilung berechnete Werte und NP_{xVP}: Werte, die nach der 1-verschobenen Poissonverteilung berechnet wurden. Die übrigen Größen sind Parameter dieser Verteilungen.

5. Die Untersuchung hat folgende Ergebnisse gebracht:

	Barock 1			Barock 2		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	72	72.67	72.96	112	109.62	111.33
2	35	33.31	33.05	24	27.96	25.36
3	9	10.18	10.11	6	4.75	5.13
4	3	2.33	2.35	1	0.61	0.97
5	0	0.52	0.53	0	0.09	0.23
Σ	119	119.01	119.00	143	143.00	143.00
	$a = 0.9168$; $k = 74.3266$; $X_2^2 = 0.24$; $p = 0.9880$; $P = 0.89$. $X_1^2 = 0.26$; $P = 0.61$.			$a = 0.5101$; $k = 2.0062$; $X_1^2 = 1.07$; $p = 0.8485$; $P = 0.30$. $X_1^2 = 0.25$; $P = 0.61$.		

Barock 1: Martin Opitz: Vom Wolffesbrunnen bey Heidelberg
Barock 2: Martin Opitz: Francisci Petrarchae

	Barock 3			Barock 4		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	108	105.86	108.01	257	253.68	
2	37	39.90	36.98	115	118.71	
3	10	10.03	10.02	42	37.03	
4	2	1.89	2.36	3	8.66	
5	1	0.34	0.65	3	1.93	
Σ	158	158.00	158.00	420	420.00	
	$a = 0.7538$; $k = 4.3507$; $X_2^2 = 0.53$; $p = 0.8720$; $P = 0.77$. $X_1^2 = 0.00$; $P = 0.99$.			$a = 0.9359$; $X_3^2 = 5.15$; $P = 0.16$.		

Barock 3: Martin Opitz: Aus dem Italienischen der edelen Poetin V.G.
Barock 4: Martin Opitz: Auf Danielis Heinsii Niederländische Poemata

	Barock 5			Barock 6		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	95	96.04	96.21	67	68.28	68.24
2	41	39.49	39.32	33	30.87	30.83
3	11	10.83	10.80	9	9.30	9.34
4	2	2.23	2.24	2	2.10	2.13
5	0	0.44	0.45	0	0.47	0.49
	149	149.00	149.00	111	111.00	111.00
	$a = 0.8224$; $k = 123.3379$; $X_2^2 = 0.23$; $p = 0.9934$; $P = 0.89$. $X_1^2 = 0.26$; $P = 0.61$.			$a = 0.9042$; $k = 175.1119$; $X_2^2 = 0.30$; $p = 0.9949$; $P = 0.86$. $X_1^2 = 0.32$; $P = 0.57$.		

Barock 5: Martin Opitz: Echo oder Wiederschall
Barock 6: Martin Opitz: Ein Gebet

	Barock 7				Barock 8		
x	n_x	NP_{xP}	NP_{xNB}	NP_{xVP}	n_x	NP_{xP}	NP_{xNB}
1	145	153.44	153.78	151.19	329	336.70	336.64
2	87	67.51	67.34	73.60	151	136.52	136.46
3	13	19.80	19.68	17.92	33	36.90	36.96
4	1	4.36	4.32	2.91	6	7.48	7.52
5	0	0.91	0.91	0.41	0	1.41	1.43
Σ	246	246.00	246.00	246.00	519	519.00	519.00
	$a = 0.8800$; $k = 874.7238$; $a = 0.4868$; $X_2^2 = 11.87$; $p = 0.9990$; $X_2^2 = 5.64$; $P = 0.003$; $X_1^2 = 11.91$; $P = 0.06$. $C = 0.05$. $P = 0.006$; $C = 0.05$.				$a = 0.8109$; $k = 464.0125$; $X_3^2 = 3.82$; p = 0.9983; $P = 0.28$. $X_2^2 = 3.87$; $P = 0.14$.		

Barock 7: Martin Opitz: Sonnet über die augen der Astree
Barock 8: Martin Opitz: An eine Jungfraw

	Barock 9			Barock 10		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	251	256.26	256.34	102	103.79	103.84
2	93	80.08	80.01	40	36.75	36.69
3	10	16.68	16.68	8	8.68	8.68
4	2	2.98	3.00	1	1.80	1.82
Σ	356	356.00	356.00	151	151.00	151.00
	$a = 0.6250$; $X_2^2 = 5.19$; $P = 0.07$.	$k = 547.2757$; $p = 0.9989$; $X_1^2 = 5.22$; $P = 0.02$; $C = 0.01$.		$a = 0.7082$; $X_2^2 = 0.71$; $P = 0.70$.	$k = 226.3046$; $p = 0.9969$; $X_1^2 = 0.74$; $P = 0.39$.	

Barock 9: Martin Opitz: Sechstine
Barock 10: Martin Opitz: Der Psalm

	Barock 11			Barock 12		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	114	118.84	118.80	62	62.83	63.00
2	62	53.08	53.08	29	27.24	27.03
3	14	15.81	15.83	7	7.87	7.86
4	2	4.28	4.30	2	2.06	2.13
Σ	192	192.00	192.00	100	100.00	100.00
	$a = 0.8933$; $X_2^2 = 3.11$; $P = 0.21$.	$k = 773.7714$; $p = 0.9988$; $X_1^2 = 3.13$; $P = 0.08$.		$a = 0.8671$; $X_2^2 = 0.22$; $P = 0.89$.	$k = 59.5726$; $p = 0.9858$; $X_1^2 = 0.26$; $P = 0.61$.	

Barock 11: Martin Opitz: An Herrn Heinrich Schützen
Barock 12: Martin Opitz: HORATII:EXEGI monumentum

	Barock 13			Barock 14	
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}
1	165	168.69	168.90	123	115.47
2	84	71.77	71.61	32	44.50
3	11	20.36	20.30	17	11.43
4	6	5.21	5.20	2	2.61
Σ	266	266.00	266.00	174	174.00
	$a = 0.8509$; $X_2^2 = 6.59$; $P = 0.04$; $C = 0.02$.	$k = 331.9038$; $p = 0.9975$; $X_1^2 = 6.62$; $P = 0.01$; $C = 0.02$.		$a = 0.7708$; $X_2^2 = 6.85$; $P = 0.03$; $C = 0.04$.	

Barock 13: Martin Opitz: Salomons hohes Lied
Barock 14: Andreas Gryphius: Eitelkeit Menschlichen Lebens

	Barock 15			Barock 16		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	92	87.45	88.94	93	91.29	92.04
2	25	32.65	30.21	28	31.20	30.05
3	12	8.13	8.28	9	7.11	7.21
4	1	1.79	2.57	1	1.41	1.72
Σ	130	130.00	130.00	131	131.00	131.00
	$a = 0.7467$; $X_2^2 = 4.21$; $P = 0.12$.	$k = 3.7463$; $p = 0.8568$; $X_1^2 = 3.62$; $P = 0.06$.		$a = 0.6835$; $X_2^2 = 0.98$; $P = 0.61$.	$k = 8.8036$; $p = 0.9334$; $X_1^2 = 0.89$; $P = 0.35$.	

Barock 15: Andreas Gryphius: Über die Geburt Jesu
Barock 16: Andreas Gryphius: Es ist alles gantz eytel

	Barock 17			Barock 18		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	92	93.68		94	94.47	94.46
2	34	30.48		32	31.30	31.21
3	6	7.86		7	6.91	6.97
4	-	-		1	1.33	1.39
Σ	132	132.00		134	134.00	134.00
	$a = 0.6507$; $X_1^2 = 0.87$; $P = 0.35$.			$a = 0.6627$; $X_2^2 = 0.10$; $P = 0.95$.		$k = 73.3575$; $p = 0.9911$; $X_1^2 = 0.12$; $P = 0.73$.

Barock 17: Andreas Gryphius: Es ist alles Eitel
Barock 18: Andreas Gryphius: Thränen in schwerer Kranckheit

	Barock 19		Barock 20	
x	n_x	NP_{xP}	n_x	NP_{xP}
1	90	85.63	76	75.46
2	48	55.33	32	32.11
3	34	23.83	8	9.11
4	3	10.23	3	2.33
Σ	175	175.00	119	119.00
	$a = 1.2923$; $X_1^2 = 1.45$; $P = 0.23$.		$a = 0.8511$; $X_2^2 = 0.34$; $P = 0.84$.	

Barock 19: Andreas Gryphius: Mitternacht
Barock 20: Andreas Gryphius: Einsamkeit

	Barock 21			Barock 22		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	96	98.06	98.02	84	85.02	85.27
2	35	29.84	29.84	35	32.11	31.90
3	4	7.11	7.14	6	8.08	8.03
4	-	-	-	2	1.80	1.81
Σ	135	135.00	135.00	127	127.00	127.00
	$a = 0.6086$; $X_1^2 = 2.29$; $P = 0.13$.		$k = 302.0266$; $p = 0.9980$; $X_0^2 = 2.31$; $C = 0.02$.	$a = 0.7552$; $X_2^2 = 0.84$; $P = 0.66$.		$k = 105.7172$; $p = 0.9930$; $X_1^2 = 0.86$; $P = 0.35$.

Barock 21: Andreas Gryphius: Als Er aus Rom geschidn
Barock 22: Andreas Gryphius: Menschliches Elende

	Barock 23			Barock 24		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	103	105.21	102.77	79	82.76	82.75
2	43	38.43	37.52	46	35.49	35.48
3	4	9.36	11.01	4	10.14	10.16
4	5	2.00	3.71	1	2.17	2.18
5	-	-	-	1	0.45	0.46
Σ	155	155.00	155.00	131	131.00	131.00
	$a = 0.7306$; $X_1^2 = 1.08$; $P = 0.30$.		$k = 3.8539$; $p = 0.8496$; $X_1^2 = 5.73$; $P = 0.02$; $C = 0.04$.	$a = 0.8576$; $X_2^2 = 7.15$; $P = 0.03$; $C = 0.05$.		$k = 656.9150$; $p = 0.9987$; $X_1^2 = 7.17$; $P = 0.01$; $C = 0.05$.

Barock 23: Andreas Gryphius: Ewige Freude des Außerwehlten
Barock 24: Andreas Gryphius: ANDREAS GRYPHIUS

	Barock 25			Barock 26	
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}
1	97	98.43	98.50	96	97.78
2	32	28.95	28.87	34	30.01
3	5	6.63	6.64	5	7.22
Σ	134	134.00	134.00	135	135.00
	$a = 0.5882$; $X_1^2 = 0.74$; $P = 0.39$.	$k = 149.9021$; $p = 0.9961$; $X_0^2 = 0.77$; $C = 0.006$.		$a = 0.6138$; $X_1^2 = 1.24$; $P = 0.27$.	

Barock 25: Andreas Gryphius: Der Tod
Barock 26: Andreas Gryphius: Das letzte Gerichte

	Barock 27			Barock 28		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	64	65.88	65.91	76	75.19	76.45
2	35	31.83	31.77	35	34.93	33.32
3	11	10.25	10.25	9	10.82	10.65
4	1	3.05	3.07	4	3.08	3.58
Σ	111	111.00	111.00	124	124.00	124.00
	$a = 0.9662$; $X_2^2 = 1.80$; $P = 0.41$.	$k = 258.2540$; $p = 0.9963$; $X_1^2 = 1.83$; $P = 0.18$.		$a = 0.9292$; $X_2^2 = 0.60$; $P = 0.74$.	$k = 8.9909$; $p = 0.9128$; $X_1^2 = 0.39$; $P = 0.53$.	

Barock 27: Andreas Gryphius: Die Hölle
Barock 28: Andreas Gryphius: Trawrklage des verwüsteten Deutschlandes

	Barock 29			Barock 30		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	91	85.42	91.24	80	81.63	81.79
2	28	33.09	26.36	36	32.04	31.93
3	6	8.54	7.91	6	8.39	8.34
4	4	1.96	3.51	2	1.96	1.95
Σ	129	129.00	129.00	124	124.00	124.00
	$a = 0.7747$; $X_2^2 = 4.07$; $P = 0.13$.	$k = 0.7957$; $p = 0.6782$; $X_1^2 = 0.63$; $P = 0.45$.		$a = 0.7851$; $X_2^2 = 1.20$; $P = 0.55$.	$k = 244.1095$; $p = 0.9968$; $X_1^2 = 1.22$; $P = 0.27$.	

Barock 29: Andreas Gryphius: Thränen des Vaterlandes
Barock 30: Andreas Gryphius: An die Sternen

	Barock 31			Barock 32		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	80	82.52	82.64	81	82.60	82.56
2	37	31.60	31.51	28	23.99	23.16
3	6	8.07	8.04	2	4.65	5.11
4	1	1.83	1.83	1	0.77	1.19
Σ	124	124.00	124.00	112	112.00	112.00
	$a = 0.7658$; $X_2^2 = 1.90$; $P = 0.39$.	$k = 323.8721$; $p = 0.9977$; $X_1^2 = 1.92$; $P = 0.17$.		$a = 0.5809$; $X_1^2 = 1.77$; $P = 0.18$.	$k = 4.5773$; $p = 0.8994$; $X_1^2 = 2.96$; $P = 0.09$.	

Barock 31: Andreas Gryphius: Morgen Sonnet
Barock 32: Andreas Gryphius: Mittag

	Barock 33			Barock 34		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	105	105.37	105.58	90	94.34	94.36
2	28	27.25	27.01	48	38.58	38.56
3	5	5.38	5.43	7	10.52	10.52
4	-	-	-	1	2.56	2.58
Σ	138	138.00	138.00	146	146.00	146.00
	$a = 0.5172$;	$k = 42.3295$;		$a = 0.8179$;	$k = 879.3219$;	
	$X_1^2 = 0.05$;	$p = 0.9882$;		$X_2^2 = 4.62$;	$p = 0.9991$;	
	$P = 0.83$.	$X_0^2 = 0.07$;		$P = 0.10$;	$X_1^2 = 4.64$;	
		$C = 0.0005$.		$C = 0.032$.	$P = 0.03$.	

Barock 33: Andreas Gryphius: Abend
Barock 34: Paul Fleming: Herrn Pauli Flemingi

	Barock 35			Barock 36		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	160	158.67	159.97	95	92.14	93.82
2	27	29.35	27.03	25	30.09	27.33
3	5	3.99	5.02	5	6.55	6.84
4	-	-	-	1	1.24	2.01
Σ	192	192.00	192.00	130	130.00	130.00
	$a = 0.3699$;	$k = 1.5235$;		$a = 0.6531$;	$k = 2.4617$;	
	$X_1^2 = 0.46$;	$p = 0.8661$;		$X_2^2 = 1.91$;	$p = 0.8317$;	
	$P = 0.51$.	$X_0^2 = 0.00$;		$P = 0.38$.	$X_1^2 = 1.40$;	
		$C = 0.000$.			$P = 0.24$.	

Barock 35: Paul Fleming: Gedancken über die Zeit
Barock 36: Paul Fleming: An Sich

	Barock 37			Barock 38		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	97	95.96	97.06	92	92.27	92.35
2	30	31.72	29.95	31	30.65	30.52
3	9	8.32	8.99	7	6.79	6.80
4	-	-	-	1	1.31	1.34
Σ	136	136.00	136.00	131	131.00	131.00
	$a = 0.6611$;	$k = 5.0310$;		$a = 0.6643$;	$k = 83.8667$;	
	$X_1^2 = 0.16$;	$p = 0.8977$;		$X_2^2 = 0.08$;	$p = 0.9922$;	
	$P = 0.69$.	$X_0^2 = 0.00$;		$P = 0.96$.	$X_1^2 = 0.10$;	
		$C = 0.00$.			$P = 0.76$.	

Barock 37: Paul Fleming: Über Gedächtniß seiner ersten Freundinn
Barock 38: Paul Fleming: Auff ihr Verbündniß

	Barock 39			Barock 40		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	90	88.73	90.31	67	72.54	72.73
2	34	34.95	32.89	48	35.44	35.35
3	8	9.18	9.12	7	11.54	11.47
4	3	2.14	2.69	1	3.48	3.45
Σ	135	135.00	135.00	123	123.00	123.00
	$a = 0.7879$;	$k = 6.0413$;		$a = 0.9771$;	$k = 769.4946$;	
	$X_2^2 = 0.55$;	$p = 0.8965$;		$X_2^2 = 8.43$;	$p = 0.9987$;	
	$P = 0.76$.	$X_1^2 = 0.22$;		$P = 0.01$;	$X_1^2 = 8.45$;	
		$P = 0.64$.		$C = 0.07$.	$P = 0.004$;	
					$C = 0.07$.	

Barock 39: Paul Fleming: Er redet die Stadt Moskau An
Barock 40: Paul Fleming: Auff die Italiaenische Weise

	Barock 41			Barock 42		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	88	87.37	88.00	87	89.60	89.86
2	16	17.15	16.00	39	33.33	33.18
3	3	2.49	3.01	6	8.27	8.19
4	-	-	-	1	1.81	1.79
Σ	107	107.00	107.00	133	133.00	133.00
	$a = 0.3926$; $k = 2.1428$; $X_1^2 = 0.19$; $p = 0.8843$; $P = 0.66$. $X_0^2 = 0.00$; $C = 0.00$.			$a = 0.7440$; $k = 379.6718$; $X_2^2 = 2.02$; $p = 0.9981$; $P = 0.37$. $X_1^2 = 2.04$; $P = 0.15$.		

Barock 41: Paul Fleming. Auff ihr Abwesen
Barock 42: Paul Fleming: An seine Thränen/ Als er von ihr verstossen war

	Barock 43			Barock 44		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	76	75.20	76.53	85	84.52	85.51
2	39	40.41	38.59	35	34.11	32.90
3	15	14.48	14.20	7	9.18	9.10
4	5	4.92	5.69	3	2.21	2.49
Σ	135	135.00	135.00	130	130.00	130.00
	$a = 1.0748$; $k = 9.6077$; $X_2^2 = 0.08$; $p = 0.9049$; $P = 0.96$. $X_1^2 = 0.13$; $P = 0.71$.			$a = 0.8070$; $k = 11.7474$; $X_2^2 = 0.84$; $p = 0.9396$; $P = 0.66$. $X_1^2 = 0.73$; $P = 0.39$.		

Barock 43: Paul Fleming: Auff den lustigen Flecken Rubar in Gilan
Barock 44: Paul Fleming: Über Herrn Martin Opitzen auf Boberfeld sein Ableben

	Barock 45			Barock 46		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	240	247.74	247.96	102	103.99	104.14
2	107	88.69	88.55	35	29.53	29.41
3	13	21.17	21.10	3	6.50	6.45
4	2	4.42	4.40	-	-	-
Σ	362	362.00	362.00	140	140.00	140.00
	$a = 0.7160$; $k = 996.6638$; $X_2^2 = 8.49$; $p = 0.9993$; $P = 0.01$; $X_1^2 = 8.51$; $C = 0.02$. $P = 0.004$; $C = 0.02$.			$a = 0.5679$; $k = 363.4014$; $X_1^2 = 2.92$; $p = 0.9984$; $P = 0.09$. $X_0^2 = 2.94$; $C = 0.02$.		

Barock 45: Philipp von Zesen: Lied von heiden-reimen
Barock 46: Philipp von Zesen: Auf di Augen seiner Liben

	Barock 47			Barock 48	
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}
1	82	81.27	82.10	106	134.32
2	25	26.11	24.84	137	90.16
3	7	6.63	7.07	32	40.34
4	-	-	-	4	13.54
5	-	-	-	2	3.64
6	-	-	-	2	1.02
Σ	114	114.00	114.00	283	283.00
	$a = 0.6427$; $k = 6.0543$; $X_1^2 = 0.08$; $p = 0.9142$; $P = 0.78$. $X_0^2 = 0.002$; $C = 0.00$.			$a = 1.3424$; $X_3^2 = 38.84$; $P = 0.00$; $C = 0.134$.	

Barock 47: Philipp von Zesen: Ein Jambisch Echonisch Sonnet
Barock 48: Philipp von Zesen: Meien-lied

	Barock 49			Barock 50		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	172	175.25	175.28	86	85.40	85.75
2	75	69.25	69.20	25	26.27	25.57
3	17	18.24	18.24	6	5.39	5.54
4	2	3.60	3.61	1	0.95	1.16
5	1	0.67	0.67	-	-	-
Σ	267	267.00	267.00	118	118.00	118.00
	$a = 0.7903$; $X_2^2 = 0.99$; $P = 0.61$.	$k = 576.2686$; $p = 0.9986$; $X_1^2 = 1.01$; $P = 0.31$.		$a = 0.6152$; $X_1^2 = 0.14$; $P = 0.71$.	$k = 10.1742$; $p = 0.9466$; $X_1^2 = 0.07$; $P = 0.79$.	

Barock 49: Philipp von Zesen: Siegeslied der himmelsflammenden Deutschen Dichtmeister
Barock 50: Philipp von Zesen: Ringel-gedichte

	Barock 51			Barock 52		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	86	86.52	86.50	89	90.31	90.47
2	34	30.09	30.03	34	31.74	31.61
3	3	6.98	7.02	7	7.44	7.40
4	1	1.21	1.24	0	1.31	1.31
5	1	0.21	0.22	1	0.23	0.23
	125	125.00	125.00	131	131.00	131.00
	$a = 0.6956$; $X_2^2 = 3.03$; $P = 0.22$.	$k = 95.4262$; $p = 0.9928$; $X_1^2 = 3.04$; $P = 0.08$.		a = 0.7028; $X_2^2 = 0.38$; $P = 0.83$.	$k = 183.3923$; $p = 0.9962$; $X_1^2 = 0.40$; $P = 0.53$.	

Barock 51: Regina von Greiffenberg: Auf meinen bestürmeten Lebens-Lauff
Barock 52: Catharina von Greiffenberg: In äusserster Widerwärtigkeit

	Barock 53			Barock 54		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	81	83.59	83.64	88	84.22	86.42
2	39	33.98	33.93	34	40.74	37.70
3	8	9.21	9.20	16	13.13	12.83
4	0	1.87	1.87	4	3.92	5.06
5	1	0.36	0.38	-	-	-
	129	129.00	129.00	142	142.00	142.00
	$a = 0.8131$; $X_2^2 = 1.65$; $P = 0.44$.	$k = 411.3148$; $p = 0.9980$; $X_1^2 = 1.67$; $P = 0.20$.		$a = 0.9673$; $X_2^2 = 1.91$; $P = 0.38$.	$k = 4.8778$; $p = 0.8516$; $X_1^2 = 1.40$; $P = 0.24$.	

Barock 53: Catharina von Greiffenberg: Auf Gottes Herrliche Wunder Regirung
Barock 54: Catharina von Greiffenberg: Verlangen nach der herrlichen Ewigkeit

	Barock 55			Barock 56		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xVP}
1	92	92.19	92.31	74	80.81	79.00
2	42	42.28	42.03	57	41.47	45.77
3	14	12.93	12.94	10	18.73	16.25
4	3	3.62	3.73	-	-	-
Σ	151	151.00	151.00	141	141.00	141.00
	$a = 0.9173$; $X_2^2 = 0.19$; $P = 0.91$.	$k = 68.6260$; $p = 0.9869$; $X_1^2 = 0.23$; $P = 0.64$.		$a = 1.0263$; $X_1^2 = 10.46$; $P = 0.001$; $C = 0.07$.	a = 0.5794; $X_1^2 = 5.47$; $P = 0.02$; $C = 0.04$.	

Barock 55: Catharina von Greiffenberg: Über den gekreuzigten JESUS
Barock 56: Catharina von Greiffenberg: Gott-lobende Frülings-Lust

	Barock 57			Barock 58		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	75	73.00	75.85	64	66.27	66.48
2	43	41.52	38.33	47	39.71	39.56
3	10	15.75	14.90	10	15.86	15.78
4	8	5.74	6.92	7	6.17	6.18
Σ	136	136.00	136.00	128	128.00	128.00
	$a = 1.1377$; $X_2^2 = 3.11$; $P = 0.21$.	k = 5.4843; $p = 0.8441$; $X_1^2 = 2.36$; $P = 0.12$.		$a = 1.1983$; $X_2^2 = 3.70$; $P = 0.16$.	$k = 180.8355$; $p = 0.9935$; $X_1^2 = 3.72$; $P = 0.05$.	

Barock 57: Catharina von Greiffenberg: Auf die Fröliche und Herrliche Auferstehung Christi
Barock 58: Catharina von Greiffenberg: Auf die Fruchtbringende Herbst-Zeit

	Barock 59			Barock 60		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	86	88.18	88.14	66	64.45	65.65
2	39	34.64	34.58	33	35.25	33.86
3	9	11.18	11.29	13	12.85	12.54
4	-	-	-	4	3.52	3.73
5	-	-	-	1	0.95	1.23
Σ	134	134.00	134.00	117	117.00	117.00
	$a = 0.7857$; $X_1^2 = 1.03$; $P = 0.31$.	$k = 118.4617$; $p = 0.9934$; $X_0^2 = 1.08$; $C = 0.008$.		$a = 1.0940$; $X_2^2 = 0.25$; $P = 0.88$.	$k = 11.9789$; $p = 0.9205$; $X_2^2 = 0.10$; $P = 0.95$.	

Barock 59: Quirinus Kuhlmann: Über den Thränen-würdigen Tod des Sohnes Gottes Jesus
Barock 60: Quirinus Kuhlmann: II. Hauptschlus des Hauptschlusses

	Barock 61			Barock 62		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	195	189.59	194.07	72	73.42	73.48
2	72	80.68	74.61	43	40.49	40.35
3	25	22.89	22.60	14	14.89	14.88
4	7	5.58	7.74	4	4.10	4.15
5	-	-	-	1	1.13	1.15
Σ	299	299.00	299.00	134	134.00	134.00
	$a = 0.8512$; $X_2^2 = 1.52$; $P = 0.47$.	$k = 4.5001$; $p = 0.8602$; $X_1^2 = 0.42$; $P = 0.52$.		$a = 1.1030$; $X_3^2 = 0.25$; $P = 0.97$.	$k = 132.5297$; $p = 0.9918$; $X_2^2 = 0.28$; $P = 0.87$.	

Barock 61: Quirinus Kuhlmann: Des 117. Kühlpsalmes I. Hauptschlus
Barock 62: Benjamin Neukirch: Über ihre unempfindligkeit

	Barock 63		Barock 64	
x	n_x	NP_{xP}	n_x	NP_{xP}
1	87	90.27	114	119.99
2	44	37.09	67	49.81
3	9	12.66	6	17.21
Σ	140	140.00	187	187.00
	$a = 0.8217$; $X_1^2 = 2.46$; $P = 0.12$.		$a = 0.8303$; $X_1^2 = 13.52$; $P = 0.0002$; $C = 0.07$.	

Barock 63: Benjamin Neukirch: An Sylvien
Barock 64 : David Schirmer: Marnia und ein Buch

	Barock 65				Barock 66		
x	n_x	NP_{xP}	NP_{xNB}	NP_{xVP}	n_x	NP_{xP}	NP_{xNB}
1	64	74.30	74.55	70.35	85	86.08	86.21
2	71	52.70	52.64	59.54	33	30.95	30.82
3	28	24.92	24.81	25.20	8	8.98	8.98
4	1	12.10	12.02	8.92	-	-	-
Σ	164	164.00	164.00	164.00	126	126.00	126.00
	$a = 1.4186$; $X_1^2 = 9.51$; $P = 0.002$; $C = 0.06$.	$k = 790.6639$; $p = 0.9982$; $X_1^2 = 18.40$; $P = 0.00$; $C = 0.11$.	$a = 0.8464$; $X_1^2 = 3.54$; $P = 0.06$.		$a = 0.7191$; $X_1^2 = 0.25$; $P = 0.61$.	$k = 126.0026$; $p = 0.9944$; $X_0^2 = 0.28$; $C = 0.002$.	

Barock 65: David Schirmer : Sie Liebet Ihn
Barock 66: David Schirmer: Seine Schwartze

	Barock 67			Barock 68		
x	n_x	NP_{xP}	NP_{xVP}	n_x	NP_{xP}	NP_{xNB}
1	81	81.10	84.00	195	200.95	200.93
2	45	35.67	36.68	99	87.77	87.74
3	4	13.24	9.32	24	25.56	25.58
4	-	-	-	2	5.58	5.60
5	-	-	-	1	1.15	1.17
Σ	130	130.00	130.00	321	321.00	321.00
	$a = 0.8797$; $X_0^2 = 0.00$; $C = 0.00$.	$a = 0.4367$; $X_1^2 = 5.03$; $P = 0.03$; $C = 0.04$.		$a = 0.8736$; $X_3^2 = 4.02$; $P = 0.26$.	$k = 607.5248$; $p = 0.9986$; $X_2^2 = 4.06$; $P = 0.13$.	

Barock 67: David Schirmer: Über seine Träume
Barock 68: Christian Hoffmann von Hoffmannswaldau: An Lauretten

	Barock 69			Barock 70		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	389	397.45	397.23	270	273.02	273.38
2	180	165.89	165.93	121	116.43	116.17
3	46	46.16	46.26	37	33.10	33.00
4	6	11.50	11.58	3	8.48	8.47
Σ	621	621.00	621.00	431	431.00	431.00
	$a = 0.8348$; $X_2^2 = 4.01$; $P = 0.13$.	$k = 848.4857$; $p = 0.9990$; $X_1^2 = 4.04$; $P = 0.04$; $C = 0.007$.		$a = 0.8529$; $X_2^2 = 4.19$; $P = 0.12$.	$k = 371.5279$; $p = 0.9977$; $X_1^2 = 4.25$; $P = 0.04$; $C = 0.01$.	

Barock 69: Christian Hoffmann von Hoffmannswaldau: Verachtung der Welt
Barock 70: Christian Hoffmann von Hoffmannswaldau: Gedancken bey Antretung des funffzigsten Jahres

	Barock 71			Barock 72		
x	n_x	NP_{xP}	NP_{xNB}	n_x	NP_{xP}	NP_{xNB}
1	66	69.50	69.63	78	77.44	78.36
2	42	34.03	33.95	32	33.44	31.98
3	7	11.11	11.07	11	9.62	9.63
4	3	3.37	3.37	2	2.51	3.03
Σ	118	118.00	118.00	123	123.00	123.00
	$a = 0.9794$; $X_2^2 = 3.60$; $P = 0.17$.	$k = 326.2303$; $p = 0.9970$; $X_1^2 = 3.63$; $P = 0.06$.		$a = 0.8635$; $X_2^2 = 0.36$; $P = 0.84$.	$k = 8.3304$; $p = 0.9125$; $X_1^2 = 0.54$; $P = 0.46$.	

Barock 71: Christian Hoffmann von Hoffmannswaldau: Sonnet. Vergänglichkeit der Schönheit
Barock 72: Christian Hoffmann von Hoffmannswaldau: Sonnet. Beschreibung vollkommener Schönheit

Barock 73

x	n_x	NP_{xP}	NP_{xNB}
1	88	88.20	88.58
2	33	31.65	31.16
3	6	7.57	7.57
4	2	1.59	1.70
Σ	129	129.00	129.00
	$a = 0.7176$;	$k = 26.6995$;	
	$X_2^2 = 0.50$;	$p = 0.9746$;	
	$P = 0.78$.	$X_1^2 = 0.50$;	
		$P = 0.48$.	

Barock 73: Christian Hoffmann von Hoffmannswaldau: Auff Ihren Schultern

6. Diese Ergebnisse sind wie folgt zu verstehen:
Die positive Poissonverteilung stellt für fast alle Texte ein geeignetes Modell dar; für wenige Texte (Gedicht 7, 56, 65) mußte die verschobene Poissonverteilung angepaßt werden.

Für die Texte 48 und 64 ließ sich keines der Modelle anpassen.
Nur in wenigen Fällen stellt die positive negative Binomialverteilung ein besseres Modell dar als die positive Poissonverteilung; meist sind die Unterschiede nur gering; es gibt in dem untersuchten Textkorpus kein einziges Gedicht, das sich nur nach der negativen Binomialverteilung, nicht aber nach der Poissonverteilung modellieren läßt.

Die positive Poissonverteilung hat sich damit für die Barockgedichte als das wesentlich geeignetere Modell erwiesen.

7. Es bietet sich an, Barockgedichte mit anderen Textsorten (Gattungen, Stilen) des Barock zu vergleichen. So ließe sich ein Überblick über die Plastizität des Deutschen in der Barockzeit gewinnen. Dabei wäre zu klären, ob man im Barock mit relativ homogenen Stilen zu rechnen hat oder nicht.

Als weitere Perspektive wäre eine Untersuchung zur Entwicklung von Wortlängenverteilungen in Gedichten als Teil der Entwicklung des Deutschen insgesamt sinnvoll. Dazu müssen allerdings für alle Phasen der Geschichte des Deutschen ähnliche Dateien erarbeitet werden, wie dies hier für das Zeitalter des Barock geschehen ist.

Literatur:

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Best, K.-H., & Altmann, G. (1996). Zur Länge der Wörter in deutschen Texten. In P. Schmidt (Hg.), *Glottometrika 15* (166-180), Trier: WVT.

Kuhr, S., & Müller, B. (1997). Zur Wortlängenhäufigkeit in Luthers Briefen. In diesem Band.

Software:

Altmann-FITTER (1994). Lüdenscheid: RAM-Verlag

Glottometrika 16, 1997, 40-54

Wortlängen in mittelhochdeutschen Texten

Karl-Heinz Best

0. Gegenstand dieser Arbeit sind mehrere mittelhochdeutsche Textgruppen: 19 Minnelieder der Zeit zwischen 1180 und 1220, 25 Texte der Sammelhandschrift Codex Karlsruhe 408, die „mit großer Wahrscheinlichkeit auf die Jahre 1430-1435" (Schmid, 1974:13) datiert werden kann, sowie die ersten 20 Kapitel des Sachsenspiegels, der nach Auskunft Thiemes in seiner Einleitung zwischen 1220 und 1235 entstanden sein soll (Sachsenspiegel: 4); eine Datierung der Vorlage fehlt leider in der hier benutzten Ausgabe und konnte bisher auch nicht aus andern Quellen bestimmt werden.

Mit dieser Untersuchung wird eine Lücke geschlossen: Bisher liegen Arbeiten zur Wortlängenverteilung in althochdeutschen Texten (Best, 1996) sowie zu allen Entwicklungsphasen des Deutschen seit frühneuhochdeutscher Zeit (vgl. Beiträge in diesem Band und Hinweise in Best & Altmann, 1996) vor. Als vorläufiges Ergebnis läßt sich daraus resümieren: Bei althochdeutschen Texten ist die Poisson-Verteilung das beste Modell; nur für einen medizinischen Text mußte die Hyperpoisson-Verteilung angepaßt werden. Für deutsche Brieftexte seit frühneuhochdeutscher Zeit ist die Hyperpoisson-Verteilung das bevorzugte Modell, das sich - mit einer einzigen Ausnahme - immer wieder erfolgreich anpassen ließ. Damit stellt sich die Frage, welche der theoretisch begründbaren (Wimmer & Altmann, 1996) und bisher bei der Anpassung an deutsche Texte bewährten Verteilungen bei mittelhochdeutschen. Texten verwendet werden kann. Der Klärung dieser Frage dient die vorliegende Arbeit.

1. An die mittelhochdeutschen Texte wurden also mit Hilfe des Altmann-Fitters (1994) die Poisson-Verteilung und die Hyperpoisson-Verteilung, jeweils in 1-verschobener Form, angepaßt. Die Ergebnisse der Anpassung der 1-verschobenen Poisson-Verteilung erwiesen sich insofern als überlegen, als in deutlich weniger Fällen Zusammenfassungen von Wortlängenklassen erforderlich waren. Aber auch die 1-verschobene Hyperpoisson-Verteilung konnte an alle bis auf einen der Texte, Kap. 1 des Sachsenspiegels (Text 45), angepaßt werden. In den Tabellen des Abschnitts 3 werden die Anpassungen der 1-verschobenen Poisson-Verteilung an die Texte dargestellt, deren Formel wie folgt lautet:

$$P_x = \frac{e^{-a}\, a^{x-1}}{(x-1)!}, \qquad x = 1,2,3,\ldots$$

2. Bei der Bearbeitung der Texte wurde immer nur der laufende Text ohne Überschrift ausgewertet. Die Bestimmung von „Wort" und „Silbe" erfolgte nach den gleichen Prinzipien wie in Best (1997: in diesem Band.). Dabei ist zu beachten, daß „Wort" strikt als orthographische Einheit bestimmt wurde. Das bedeutet, daß z.B. enklitische Pronomina nicht als Wörter berücksichtigt wurden, während getrennt geschriebene Präfixe als Wörter galten. Bei verschiedener Schreibung desselben Wortes wurde entsprechend verfahren. In Zweifelsfällen wurden entsprechende Handbücher wie Paul u.a. (1982) zu Rate gezogen.

3. Die Ergebnisse finden sich in den Tabellen. Die Anpassung der 1-verschobenen Poisson-Verteilung an die Daten der Texte gilt als erfolgreich, wenn $P \geq 0.05$ (P = Wahrscheinlichkeit des gegebenen oder eines noch extremeren Chiquadrat-Wertes). Als noch akzeptabel gilt es, wenn P sich im Bereich von $0.01 \leq P < 0.05$ bewegt. Falls P mangels Freiheitsgraden nicht bestimmt werden kann, wird die Anpassung mittels des Diskrepanzkoeffizienten $C = X^2/N$ bewertet, für den die Bedingung $C \leq 0.02$ erfüllt sein muß.

Die Ergebnisse:

Dabei bedeutet:

x - Wortlänge (in Silben);
n_x - Zahl der Wörter mit x Silben im Text;
NP_x - Zahl der aufgrund der 1-verschobenen Poisson-Verteilung berechneten Wörter mit x Silben;
X^2 - Chiquadrat;
FG - Freiheitsgrade; und
a - Parameter der Poisson-Verteilung.

a. Minnelieder

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	178	170.20	131	135.37	116	115.68
2	72	85.33	77	65.73	47	47.88
3	22	21.39	9	15.95	11	9.91
4	6	3.57	2	2.58	1	
5	2	0.44	1	0.37		1.53
6	1	0.07				
	$a = 0.5013$; $X_1^2 = 3.664$; $P = 0.06$.		$a = 0.4856$; $X_2^2 = 5.109$; $P = 0.08$.		a =0.4139; $X_2^2 = 0.316$; $P = 0.85$.	

Text 1: Walther von der Vogelweide, Der Minne Gewalt. In: Deutscher Minnesang. Einführung ... von Friedrich Neumann. Stuttgart: Reclam 1995. 64-66
Text 2: ders., Maiwunder. Quelle: wie Text 1. 66-70
Text 3: ders., Zweisamkeit. Quelle: wie Text 1. 70-72

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	137	136.88	83	78.20	119	118.52
2	86	73.39	40	45.54	57	57.17
3	10	19.67	12	13.26	13	13.79
4	1	4.06	5	3.00	3	2.52
	$a = 0.5362$; $FG = 0$.		$a = 0.5823$; $X_2^2 = 2.427$; $P = 0.30$.		a =0.4824; $X_2^2 = 0.142$; $P = 0.93$.	

Text 4: ders., Traumliebe. Quelle: wie Text 1. 72-74
Text 5: ders., Unter der Linde. Quelle: wie Text 1. 74-76
Text 6: ders., Liebe macht schön. Quelle: wie Text 1. 78-80

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	118	121.87	54	155.55	42	41.76
2	73	65.14	72	69.38	14	14.42
3	10	17.41	15	15.47	3	2.82
4	7	3.58	2	2.60		
	$a = 0.5345$; $X_1^2 = 1.824$; $P = 0.18$.		$a = 0.4461$; $X_2^2 = 0.260$; $P = 0.88$.		a.= 0.3454; $X_1^2 = 0.027$; $P = 0.87$.	

Text 7: ders., Das bessere Spiel. Quelle: wie Text 1. 80-82
Text 8: ders., Männerwille, Frauensitte. Quelle: wie Text 1. 82-84
Text 9: Albrecht von Johannsdorf, Wunsch vor der Kreuzfahrt. Quelle: wie Text 1. 60-62

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	116	119.20	165	166.99	97	97.33
2	58	51.09	70	66.04	36	35.37
3	9	12.71	13	14.97	6	6.43
4					1	0.87
	$a = 0.4287$; $X_1^2 = 2.096$; $P = 0.15$.		$a = 0.3955$; $X_1^2 = 0.519$; $P = 0.47$.		$a = 0.3635$; $X_1^2 = 0.023$; $P = 0.88$.	

Text 10: ders., Minne und Treue. Quelle: wie Text 1. 62-64
Text 11: Hartmann von Aue, Absage und Rückkehr. Quelle: wie Text 1. 42-44
Text 12: ders., Überhöhte Minne. Quelle: wie Text 1. 44-46

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	107	110.12	136	139.30	62	63.79
2	52	44.52	66	59.15	30	25.40
3	3	9.00	10	12.55	3	5.81
4	3	1.36	1	2.00		
	$a = 0.4043$; $X_1^2 = 3.170$; $P = 0.08$.		$a = 0.4246$; $X_2^2 = 1.880$; $P = 0.39$.		$a = 0.3982$; $X_1^2 = 2.235$; $P = 0.13$.	

Text 13: Heinrich von Morungen, Selige Tage. Quelle: wie Text 1. 52-54
Text 14: ders., Minnezauber. Quelle: wie Text 1. 54-56
Text 15: ders., Nein, ja! Quelle: wie Text 1. 56-58

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	87	89.11	44	44.89	147	142.63
2	43	38.32	29	25.37	53	60.50
3	6	8.24	4	7.16	11	12.83
4	1	1.33	2	1.58	7	2.04
	$a = 0.4301$; $X_2^2 = 1.308$; $P = 0.52$.		$a = 0.5651$; $X_2^2 = 2.059$; $P = 0.36$.		$a = 0.4242$; $X_1^2 = 1.729$; $P = 0.19$.	

Text 16: ders., Auf mein Grab. Quelle: wie Text 1. 58-60
Text 17: ders., Seelenminne. Quelle: wie Text 1. 60
Text 18: Reinmar (von Hagenau), Die Klage der Witwe. Quelle: wie Text 1. 46-48

	Text 19	
x	n_x	NP_x
1	283	277.13
2	101	109.21
3	22	21.51
4	3	2.82
5	2	0.33
	$a = 0.3941$; $X_2^2 = 1.872$; $P = 0.39$.	

Text 19: ders., Der Minne bleiche Farbe. Quelle: wie Text 1. 48-52

b. Texte aus dem Codex Karlsruhe 408

	Text 20		Text 21		Text 22	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	116	112.77	164	163.86	135	134.93
2	51	56.42	45	45.26	74	61.60
3	11	14.11	7	6.88	4	16.47
4	4	2.35				
5	4	0.35				
	$a = 0.5004$; $X_1^2 = 0.903$; $P = 0.34$.		$a = 0.2762$; $X_1^2 = 0.004$; $P = 0.95$.		$a = 0.4565$; $FG = 0$.	

Text 20: Der spunczeniererin gebet. Quelle: Codex Karlsruhe 408. Bearb. v. Ursula Schmid. 188
Text 21: Der hunt mit dem bein. Quelle: wie Text 20. 208
Text 22: Von der kriebsein. Quelle: wie Text 20. 264-5

	Text 23		Text 24		Text 25	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	236	232.61	160	158.73	223	225.26
2	68	74.19	48	51.80	91	86.69
3	15	11.83	10	8.45	15	16.68
4	0	1.25	1	0.92	2	2.37
5	1	0.12	1	0.10		
	a =0.3189; X_2^2 = 1.512; P = 0.47.		a = 0.3264; X_1^2 = 0.975; P = 0.32.		a = 0.3849; X_2^2 = 0.462; P = 0.79.	

Text 23: Das opffer kalb. Quelle: wie Text 20. 266-7
Text 24: Von dem wolff vnd hund. Quelle: wie Text 20. 294
Text 25: Von dem storg der frosch got. Quelle: wie Text 20. 295-6

	Text 26		Text 27		Text 28	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	136	137.57	138	138.75	170	173.05
2	54	50.78	56	52.79	70	62.96
3	9	10.65	7	10.04	9	12.99
4			2	1.42		
	a = 0.3692; X_1^2 = 0.475; P = 0.49.		a = 0.3805; X_2^2 = 1.373; P = 0.50.		a = 0.3638; X_1^2 = 2.060; P = 0.15.	

Text 26: Von der swalben. Quelle: wie Text 20. 297
Text 27: Von dem weyhen vnd seiner muter. Quelle: wie Text 20. 300
Text 28: Von dem lewen dem ochsen dem eßel/ vnd dem swein. Quelle: wie Text 20. 301-2

	Text 29		Text 30		Text 31	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	149	151.48	133	130.68	104	106.01
2	56	50.20	40	45.42	44	39.58
3	6	9.32	11	7.89	6	8.41
4			1	1.01		
	a = 0.3314; X_1^2 = 1.892; P = 0.17.		a = 0.3476; X_1^2 = 1.773; P = 0.18.		a =0.3734 X_1^2 = 1.218 P = 0.27.	

Text 29: Von dem lewen vnd der meus. Quelle: wie Text 20. 303
Text 30. Von dem Reyger. Quelle: wie Text 20. 307
Text 31: Von dem grillen und der emeyß. Quelle: wie Text 20. 309

	Text 32		Text 33		Text 34	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	84	84.61	297	302.01	169	167.53
2	28	26.69	116	104.59	43	45.64
3	4	4.70	13	18.11	6	6.21
4			1	2.29	2	0.62
	a = 0.3155; X_1^2 = 0.170; P = 0.68.		a = 0.3463; X_2^2 = 3.492; P = 0.17.		a = 0.2724; X_1^2 = 0.369; P = 0.54.	

Text 32: Wie man vmb das krenczlin biten soll. Quelle: wie Text 20. 519
Text 33: Von dem gutten hannen.Quelle: wie Text 20. 304-5
Text 34: Von dem lewen wolff vnd dem fuchs. Quelle: wie Text 20. 306

	Text 35		Text 36		Text 37	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	156	150.36	300	302.11	285	287.56
2	39	48.06	109	105.23	122	117.70
3	1	7.68	18	18.32	24	24.08
4	1	0.90	1	2.34	2	3.66
	a =0.3196; X_1^2 =3.293; P =0.07.		a = 0.3483; X_2^2 = 0.912; P = 0.63.		a = 0.4093; X_2^2 = 0.928; P = 0.63.	

Text 35: Von dem fuchs vnd der kaczen. Quelle: wie Text 20. 308
Text 36: Der keßdiep. Quelle: wie Text 20. 259-60
Text 37: Das eselspiel. Quelle: wie Text 20. 261-3

	Text 38		Text 39		Text 40	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	146	148.50	178	176.02	151	146.09
2	61	55.63	58	61.75	59	65.66
3	9	11.87	13	10.83	15	14.75
4			1	1.40	4	2.50
	a =0.3746; X_1^2 =1.247; P =0.26.		a = 0.3508; X_2^2 = 0.792; P = 0.67.		a = 0.4495; X_2^2 1.771; P = 0.41.	

Text 38: Von der snecken. Quelle: wie Text 20. 310
Text 39: Von des schuchsters kaczen. Quelle: wie Text 20. 311-2
Text 40: Von dem engel michahel. Quelle: wie Text 20. 463-4

	Text 41		Text 42		Text 43	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	290	279.75	271	273.62	440	440.06
2	92	108.97	96	86.96	182	164.13
3	27	21.22	7	13.81	16	30.61
4	4	3.06	2	1.61	1	4.20
	a = 0.3895; X_2^2 = 4.889; P = 0.09.		a = 0.3178; X_2^2 = 4.436; P = 0.11.		a = 0.3730; FG = 0;	

Text 41: Wie got den menschen macht. Quelle: wie Text 20. 465-7
Text 42: Von der mynne krafft. Quelle: wie Text 20. 517-8
Text 43: Von dem knecht herolt. Quelle: wie Text 20. 209-11

	Text 44	
x	n_x	NP_x
1	323	322.97
2	131	117.01
3	10	24.02
4		
5		
	a =0.3623 FG = 0	

Text 44: Der esel mit des lewen haút. Quelle: wie Text 20. 212-213

c. Texte aus dem Sachsenspiegel

(Anm.: In einigen Texten des Sachsenspiegels werden Ziffern zur Textgliederung verwendet; diese wurden nicht mitgezählt.)

	Text 45		Text 46		Text 47	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	45	44.98	95	95.01	215	212.07
2	20	31.19	82	76.70	145	153.81
3	22	10.81	23	30.95	67	55.77
4	3	3.02	13	10.34	11	16.35
	a =0.6936; FG = 0.		a = 0.8073; X_2^2 = 3.105; P = 0.21.		a= 0.7253; X_2^2 = 4.544; P = 0.10.	

Text 45: Sachsenspiegel, Landrecht, 1. Buch, Kap. 1, S. 20
Text 46: dass., Kap. 2, S. 20f.
Text 47: dass., Kap. 3, S. 21f.

	Text 48		Text 49		Text 50	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	48	45.99	110	117.14	90	89.63
2	26	29.84	93	79.03	50	50.80
3	13	9.68	15	26.66	15	14.40
4	0	2.09	10	5.99	2	2.72
5	1	0.40	2	1.18	1	0.45
	a = 0.6489; X_2^2 = 2.605; P = 0.27.		a = 0.6746; X_1^2 = 4.279; P = 0.04.		a = 0.5668; X_2^2 = 0.047; P = 0.98.	

Text 48: dass., Kap. 4, S. 23
Text 49: dass., Kap. 5, S. 23f.
Text 50: dass., Kap. 6, S. 24

	Text 51		Text 52		Text 53	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	35	34.52	62	62.39	189	190.85
2	17	17.90	42	41.83	116	114.09
3	5	4.64	15	14.02	38	34.10
4	1	0.94	3	3.76	4	7.96
	a = 0.5187; X_1^2 = 0.087; P = 0.77.		a = 0.6705; X_2^2 = 0.217; P = 0.90.		a = 0.5978; X_2^2 = 2.455; P = 0.29.	

Text 51: dass., Kap. 7, S. 24f.
Text 52: dass., Kap. 8, S. 25
Text 53: dass., Kap. 9, S. 25f.

	Text 54		Text 55		Text 56	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	33	33.98	30	29.93	55	55.00
2	31	23.57	24	21.30	34	33.96
3	1	8.17	4	7.58	13	13.04
4	3	2.28	3	2.19		
	a = 0.6937; X_2^2 = 8.902; P = 0.01.		a = 0.7118; X_2^2 = 2.350; P = 0.31.		a = 0.6176; FG = 0; P = 0.99.	

Text 54: dass., Kap. 10, S. 26
Text 55: dass., Kap. 11, S. 27
Text 56: dass., Kap. 12, S. 27

	Text 57		Text 58		Text 59	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	70	67.54	56	54.29	61	55.77
2	42	45.76	32	35.29	30	36.85
3	16	15.50	15	11.46	12	12.17
4	3	3.50	1	2.96	5	3.21
5	2	0.70				
	$a = 0.6775$; $X_2^2 = 0.574$; $P = 0.75$.		$a = 0.6500$; $X_2^2 = 2.734$; $P = 0.25$.		$a = 0.6608$; $X_2^2 = 2.793$; $P = 0.25$.	

Text 57: dass., Kap. 13, S. 27f.
Text 58: dass., Kap. 14, S. 28
Text 59: dass., Kap. 15, S. 28f.

	Text 60		Text 61		Text 62	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	38	37.04	83	85.85	65	61.12
2	21	21.95	56	50.16	32	39.16
3	6	6.50	12	14.65	16	12.54
4	2	1.51	2	2.85	3	3.18
5			1	0.49		
	$a = 0.5924$; $X_2^2 = 0.272$; $P = 0.87$.		$a = 0.5843$; $X_2^2 = 1.287$; $P = 0.53$.		$a = 0.6407$; $X_2^2 = 2.516$; P 0 0.28.	

Text 60: dass., Kap. 16, S. 29
Text 61: dass., Kap. 17, S. 29f.
Text 62: dass., Kap. 18. S. 30

	Text 63		Text 64	
x	n_x	NP_x	n_x	NP_x
1	44	44.81	163	172.10
2	37	36.41	135	114.11
3	17	14.79	26	37.83
4	3	4.99	9	8.36
5			1	1.60
	$a = 0.8125$; $X_2^2 = 1.136$; $P = 0.57$.		$a = 0.6630$; $X_3^2 = 8.274$; $P = 0.04$.	

Text 63: dass., Kap. 19, S. 30f.
Text 64: dass., Kap. 20, S. 31f.

4. Ergebnis: Es hat sich gezeigt, daß die 1-verschobene Poisson-Verteilung an alle 64 mittelhochdeutschen Texte angepaßt werden kann. Es erstaunt, daß auch Text 20, ein deutsch-lateinischer Mischtext, mit gutem Ergebnis modelliert werden konnte.

Wie eingangs erwähnt, läßt sich auch die 1-verschobene Hyperpoisson-Verteilung mit Erfolg anpassen; sie versagt lediglich bei einem der Texte (Text 45). Bei der Anwendung der Hyperpoisson-Verteilung muß allerdings wesentlich häufiger auf den Diskrepanzkoeffizienten C als Prüfgröße zurückgegriffen werden, da die mittelhochdeutschen Texte oft nur 3- und 4-silbige Wörter enthalten. In solchen Fällen lassen sich Modelle, die wie die Poisson-Verteilung nur einen Parameter aufweisen, besser anpassen als z.B. die Hyperpoisson-Verteilung mit ihren zwei Parametern. Es könnte sich herausstellen, daß die Hyperpoisson-Verteilung auch für mittelhochdeutsche. Texte ein gleichwertiges Modell darstellt, wenn man Texte mit größeren Wortlängen bearbeitet.

5. Mit dieser Bearbeitung mittelhochdeutscher. Texte liegen erste Ergebnisse zur Wortlängenverteilung für alle schriftlich überlieferten Phasen des Deutschen vor. Dabei handelt es sich überwiegend um literarische Texte und Briefe. In den frühen Phasen scheint die Poisson-Verteilung das geeignetste Modell zu sein; ab frühneuhochdeutscher Zeit ergibt die Hyperpoisson-Verteilung bisher die besten Ergebnisse.

Es bleibt zu untersuchen, ob dies nur für die bisher bearbeiteten Textsorten gilt, oder ob bei Berücksichtigung anderer, z.B. fachsprachlicher oder wissenschaftlicher Texte nicht doch weitere Modelle in Betracht gezogen werden können oder gar müssen. Es gibt deutliche Hinweise, daß für gegenwärtiges Deutsch z.B. die positive negative Binomialverteilung bei Pressetexten (Best, 1997) bzw.

die positive Singh-Poisson-Verteilung bei naturwissenschaftlichen Texten (Behrmann, 1997) gute Ergebnisse zeitigen, beides Verteilungen, die sich bei älteren deutschen Texten als weniger oder gar nicht geeignet erwiesen.

Quellentexte

Codex Karlsruhe 408. Bearbeitet von Ursula Schmid. Bern/ München: Francke 1974 [Schmid 1974]

Deutscher Minnesang. Einführung sowie Auswahl und Ausgabe der mittelhochdeutschen Texte von Friedrich Neumann. Nachdichtung von Kurt Erich Meurer. Suttgart: Reclam 1995

Sachsenspiegel. (Landrecht). Hg. v. Cl. Frhr. von Schwerin. Eingel. von Hans Thieme. Stuttgart: Reclam 1977

Literatur

Behrmann, G. (1997). Die Wortlängenhäufigkeiten von deutschsprachigen naturwissenschaftlichen Publikationen. In diesem Band.

Best, K.-H. (1996). Zur Bedeutung von Wortlängen, am Beispiel althochdeutscher Texte. *Papiere zur Linguistik.* Im Druck.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In diesem Band.

Best, K.-H., & Altmann, G. (1996). Project Report. *Journal of Quantitative Linguistics*, 3,1, 85-88.

Paul, H., Moser, H., Schröbler, I., & Grosse, S. (1982). *Mittelhochdeutsche Grammatik.* 22., durchges. Auflage. Tübingen: Niemeyer.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

Software

Altmann-FITTER (1994). Lüdenscheid: RAM-Verlag.

Glottometrika 16, 1997, 55-62

Zur Wortlängenhäufigkeit in Luthers Briefen

Saskia Kuhr, Barbara Müller

0. Bei der Untersuchung der Wortlängenhäufigkeiten im Deutschen hat sich bisher gezeigt, daß die positive (null-gestutzte) negative Binomialverteilung ein geeignetes Modell für etliche Textsorten des Deutschen der Gegenwart und der nahen Vergangenheit ist. Sie konnte u.a. an literarische Prosa (Best & Zhu, 1994), deutschsprachige Pressetexte (Best, 1997) und deutsche Briefe (Ammermann, 1997; Bartels & Strehlow, 1997) angepaßt werden. Nur Texte für Kinder bis zu 12 Jahren zeigten Abweichungen (Laass, 1996); auch Pressetexte mit deutlich fachsprachlichem Einschlag ergaben einige Abweichungen (vgl. dazu Riedemann, 1994). Insgesamt gesehen wirkt das Deutsche hinsichtlich der Wortlängenhäufigkeiten jedoch relativ homogen, vergleicht man es mit den Ergebnissen, die bei anderen Sprachen erarbeitet wurden (vgl. Altmann, Erat & Hřebíček, 1996).

1. Noch fast ganz unbekannt ist, wie und ob sich die Wortlängenhäufigkeit in der Geschichte des Deutschen entwickelt hat. So bot es sich an, an den Beginn der Entwicklung des Neuhochdeutschen zurückzugehen und nach geeigneten Texten aus frühneuhochdeutscher Zeit zu suchen.

Bei der Suche nach einem geeigneten frühneuhochdeutschen Autoren fiel die Wahl auf Luther, da er bekanntlich, im Gegensatz zu vielen gelehrten Zeitgenossen, einen Großteil seiner Schriften in deutscher Sprache abgefaßt hat, um sie auch Laien zugänglich zu machen. Viele der Schriften Luthers sind jedoch recht umfangreich, so daß sie sich nicht für die Analyse eignen würden. Die zu analysierenden Texte sollten „in einem Guß" geschrieben worden sein, um Homogenität zu gewährleisten. Unsere Textauswahl beschränkt sich daher auf Briefe Luthers, da man bei Briefen davon ausgehen kann, daß sie an einem Stück geschrieben wurden. Eine Ausnahme stellt die Vorrede zum Kleinen Katechismus dar, die aufgrund ihrer Kürze (1441 Wörter insgesamt) ebenfalls geeignet erschien und daher ebenfalls zur Analyse herangezogen wurde.

Innerhalb der Korrespondenz Luthers ist im Verhältnis lateinischer und deutscher Texte eine Entwicklung feststellbar. „Rund 3/5 seiner Briefe sind latei-

nisch; 1517 erscheint erstmals ein deutscher Brief neben 20 lateinischen. Doch erst in den 20er Jahren zieht L(uther) zu seiner Korrespondenz in großem Ausmaß auch die Volkssprache heran" (Wolf, 1980:148). Die Verwendung der beiden Sprachen hängt vom jeweiligen Briefpartner ab. Da Luther an Gelehrte und gebildete Geistliche meist lateinisch schrieb, richteten wir unser Augenmerk vor allem auf Briefe an Fürsten, städtische Würdenträger und Privatpersonen. Die Briefe Luthers an seine Frau Katharina sind ausnahmslos auf Deutsch geschrieben. Doch auch innerhalb deutscher Texte sind lateinische Einflechtungen zu finden. Wir haben die lateinischen Worte und Wendungen fortlaufend mitgezählt, da sie nie die Länge eines Satzes überschreiten. Die Auswertungen haben gezeigt, daß sich diese Vorgehensweise nicht negativ ausgewirkt hat. Ein weiterer Vorteil der Beschränkung auf Briefe ist, daß sie ein „gesprächsnahes Mittel" darstellen und Luther bei „wachsender Abkehr von der humanistischen Epistolographie auf rhetorische Eleganz verzichtet, dafür um so mehr den Regeln der Dialektik folgt" (Wolf, 1980:148).

2. Bei der Bearbeitung der Texte wurden folgende Prinzipien eingehalten: Alle Briefe wurden vollständig, d.h. mit Anrede, Ort, Datum, laufendem Text, Unterschrift und ggfs. Postskriptum ausgewertet. Diese Erhebungsprinzipien wurden nicht nachträglich revidiert, da sie sich bei der Auswertung nicht störend bemerkbar gemacht haben.

Auch bei diesen frühneuhochdeutschen Texten wurde „Wort" orthographisch definiert. Sonderfälle, wie z.B. die Kontraktion „hettestu", wurden dementsprechend als ein Wort behandelt. Nullsilbige Wörter wurden nicht ausgenommen, so daß apostrophiertes „'s" dem jeweils vorausgehenden Wort zugerechnet wurde (z.B. „tut's" wurde als ein einsilbiges Wort gewertet). Die Zahl der Silben wurde aufgrund der Operationalisierung anhand der Zahl der Vokale bzw. Diphthonge im Wort bestimmt. Dabei ist der Schreibung <y> für [i] oder <w> für [u] Rechnung zu tragen (vgl. zur Graphematik Luthers: Wolf, 1980:31ff). Bei Unsicherheit in Bezug auf die Aussprache wurde einheitlich buchstabengetreu gezählt; so wurde z.B. „ewr" als einsilbiges Wort gewertet.

Zahlwörter wurden entsprechend der Orthographie als ein Wort gewertet. Da es sich dabei meist nur um das am Ende des Briefes stehende Datum handelt, hat sich die Mitzählung nicht negativ auf die Berechnungen ausgewirkt. Die Abkürzungen wurden wie gelesen mitgezählt, z.B. „G. und fried" = Gnad und fried. Bei der Entschlüsselung wurde auf die Inselausgabe (Bornkamm & Ebeling, 1982) zurückgegriffen, die die meisten Abkürzungen ausschreibt und die restlichen am Ende des Briefbandes aufführt.

Die Silbenzählung ergab, daß die analysierten Texte hauptsächlich ein- bis dreisilbige Wörter enthalten, und zwar etwas weniger zweisilbige als einsilbige Wörter, sehr viel weniger drei- als zweisilbige Wörter und nur vereinzelt viersilbige. Dieses mag vor allem an der Flexion der Wörter im 16. Jh. liegen, da En-

dungen häufig wegfallen und die Wörter daher meist kürzer sind. So sind zahlreiche Substantive im Plural bei Luther einsilbig, die heute zweisilbig sind; Neutra im Nominativ und Akkusativ Plural sind häufig endungslos, z.B. „die wort" statt „die Worte". Auch Adjektive im Nominativ Singular Maskulinum, Femininum und Neutrum sowie im Akkusativ Singular sind in der Regel endungslos, wie z.B. „ein frei Konzil". Das auslautende [e] bei Adjektiven wird häufig apokopiert („unser Kirche"). Es finden sich bei Luther aber auch zum Teil schon die moderneren Formen, wie z.B. „Gnade" neben „Gnad". „In der Morphologie erweist sich L(uther) als Repräsentant einer sprachgeschichtlichen Übergangsstufe; einerseits ist er noch zahlreichen älteren Formen verhaftet, andererseits schlägt er im Laufe seines Schaffens allerdings auch schon deutliche Schritte in Richtung auf den nhd. Formenstand ein, obschon kaum konsequent" (Wolf, 1980:33). Bei fünfsilbigen Wörtern, die sehr selten auftreten, handelt es sich um lateinische oder griechische Fremdwörter oder um ehrerbietige Anreden in offiziellen Briefen („unterthäniglich"). Bei sechs- bis achtsilbigen Wörtern handelt es sich in den meisten Fällen um Zahlwörter.

3. Die Daten der Briefe und der Vorrede zum Kleinen Katechismus wurden auf mögliche Modelle hin überprüft. Es zeigt sich, daß die 1-verschobene Hyperpoisson-Verteilung an alle Texte angepaßt werden konnte, deren Formel lautet:

$$P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1\,(1;\,b;\,a)}, \quad x = 1,2,3,...$$

$$c^{(0)} = 1$$

$$c^{(x)} = c\,(c+1)\,...\,(c+x-1)$$

$${}_1F_1\,(1;b;a) = \sum_{j=0}^{\infty} \frac{a^j}{b^{(j)}}$$

Die Anpassung des Modells an die einzelnen Texte gilt als zufriedenstellend, wenn $P \geq 0.05$ bzw., besonders bei längeren Texten, $C \leq 0.02$. Dabei ist P die Überschreitungswahrscheinlichkeit des Chiquadrats, $C = X^2/N$ der Diskrepanzkoeffizient. Die Tabellen in Abschnitt 4 zeigen nun, daß die Hyperpoisson-Verteilung in allen Fällen an die Texte angepaßt werden konnte. In zwei Texten erwies sich P als nicht zufriedenstellend (wohl aber noch akzeptabel); dafür genügt C bei diesen Texten den genannten Bedingungen.

Außer diesen Prüfgrößen enthalten die Tabellen folgende Informationen: x ist die Wortlänge in Silben, n_x ihre beobachtete Häufigkeit im Text, NP_x die aufgrund der Hyperpoissonverteilung errechnete Häufigkeit, a, b sind die Parameter der Verteilung; X^2 ist das Chiquadrat.

4. Im Einzelnen haben die Untersuchungen folgende Ergebnisse erbracht:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	108	106.46	272	272.78	156	156.56
2	102	100.54	181	180.55	101	101.36
3	24	28.58	54	51.70	24	21.93
4	5	4.78	7	9.44	1	2.85
5	0	0.56	1	1.26	0	0.26
6	1	0.05	1	0.18	1	0.04
7	0	0.00	-	-	-	-
8	1	0.03	-	-	-	-
	$a = 0.41$; $b = 0.43$; $X_1^2 = 1.25$; $P = 0.26$.		$a = 0.50$; $b = 0.76$; $X_2^2 = 0.98$; $P = 0.61$.		$a = 0.33$; $b = 0.50$; $X_1^2 = 0.61$; $P = 0.44$.	

Text 1: Brief Luthers an Katharina Luther vom 4. 10. 1529.
Text 2: Brief Luthers an seine „Tischgesellen" in Wittenberg vom 26.4. 1530.
Text 3: Brief Luthers an Katharina Luther vom 5.6.1530.

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	249	246.19	322	322.32	216	210.60
2	163	161.16	186	186.18	164	173.66
3	35	41.94	45	44.12	68	55.15
4	7	6.81	6	6.57	3	10.84
5	2	0.80	0	0.71	1	1.75
6	0	0.07	0	0.06	-	-
7	0	0.00	0	0.00	-	-
8	1	0.03	1	0.04	-	-
	$a = 0.43$; $b = 0.66$; $X_1^2 = 1.89$; $P = 0.17$.		$a = 0.40$; $b = 0.70$; $X_1^2 = 0.04$; $P = 0.85$.		$a = 0.52$; $b = 0.63$; $X_2^2 = 9.66$; $P = 0.01$; $C = 0.02$.	

Text 4: Brief Luthers an Katharina Luther vom 27.2.1532.
Text 5: Brief Luthers an Katharina Luther vom 7.2.1546.
Text 6: Brief Luthers an Kurfürst Friedrich den Weisen vom 6.11.1517.

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	113	112.79	154	153.85	426	425.12
2	73	77.69	100	99.90	235	245.37
3	39	30.25	28	28.33	113	95.87
4	6	10.27	2	5.13	21	28.31
5	-	-	4	0.79	7	6.72
6	-	-	-	-	0	1.33
7	-	-	-	-	0	0.22
8	-	-	-	-	1	0.06
	$a = 0.90$; $b = 1.30$; $X_1^2 = 4.58$; $P = 0.03$; $C = 0.02$.		$a = 0.50$; $b = 0.78$; $X_1^2 = 0.01$; $P = 0.94$.		$a = 1.21$; $b = 2.10$; $X_3^2 = 5.63$; $P = 0.13$.	

Text 7: Brief Luthers an Kurfürst Friedrich den Weisen vom 15.5.1519.
Text 8: Brief Luthers an Kurfürst Friedrich vom 24.2.1522.
Text 9: Brief Luthers an Kurfürst Johann vom 31.10. 1525.

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	279	279.23	791	789.96	102	102.41
2	149	149.12	491	484.16	58	58.23
3	54	51.92	126	137.53	14	12.52
4	11	13.41	28	25.42	0	1.66
5	3	2.75	5	3.93	0	0.15
6	0	0.46	-	-	0	0.01
7	0	0.06	-	-	0	0.00
8	1	0.05	-	-	1	0.02
	$a = 1.00$; $b = 1.87$; $X_2^2 = 0.67$; $P = 0.72$.		$a = 0.53$; $b = 0.86$; $X_2^2 = 1.64$; $P = 0.44$.		$a = 0.35$; $b = 0.61$; $X_1^2 = 0.56$; $P = 0.46$.	

Text 10: Brief Luthers an den Landgrafen Philipp von Hessen vom 7.1.1527.
Text 11: Vorrede zum Kleinen Katechismus.
Text 12: Brief Luthers an Katharina Luther vom 29.7.1534.

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	184	185.42	220	219.55	108	106.82
2	104	104.80	132	126.89	79	78.13
3	34	29.41	40	42.44	18	21.08
4	2	5.49	5	9.99	3	3.49
5	1	0.76	3	1.81	1	0.41
6	0	0.08	1	0.32	1	0.07
7	0	0.01	-	-	-	-
8	0	0.00	-	-	-	-
9	1	0.04	-	-	-	-
	$a = 0.56$; $b = 0.99$; $X_1^2 = 1.60$; $P = 0.21$.		$a = 0.79$; $b = 1.37$; $X_2^2 = 4.51$; $P = 0.10$.		$a = 0.43$; $b = 0.58$; $X_1^2 = 0.76$; $P = 0.38$.	

Text 13: Brief Luthers an Katharina Luther vom 27.2.1537.
Text 14: Brief Luthers an Katharina Luther vom 2.7.1540.
Text 15: Brief Luthers an Katharina Luther vom 16.7.1540.

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	124	124.64	71	71.83	222	221.84
2	102	102.52	29	29.34	120	119.91
3	30	25.62	11	8.50	38	38.00
4	0	3.77	0	1.91	8	8.52
5	0	0.39	0	0.35	1	1.47
6	1	0.06	0	0.05	0	0.20
7	-	-	0	0.00	0	0.02
8	-	-	1	0.02	1	0.04
	$a = 0.36$; $b = 0.44$; $X_1^2 = 3.20$; $P = 0.07$.		$a = 1.00$; $b = 2.44$; $X_1^2 = 1.50$; $P = 0.23$.		$a = 0.77$; $b = 1.42$; $X_2^2 = 0.08$; $P = 0.96$.	

Text 16: Brief Luthers an Katharina Luther vom 26.7.1540.
Text 17: Brief Luthers an Katharina Luther vom 18.9.1541.
Text 18: Brief Luthers an Katharina Luther vom 28.7.1545.

	Text 19		Text 20		Text 21	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	132	131.56	178	180.29	159	159.39
2	78	77.74	91	87.20	83	83.20
3	27	26.75	27	29.17	30	28.42
4	5	6.49	8	7.46	6	7.21
5	1	1.21	0	1.54	0	1.45
6	0	0.18	1	0.26	1	0.24
7	0	0.02	0	0.04	0	0.03
8	1	0.05	1	0.04	1	0.06
	$a = 0.82$; $b = 1.40$; $X_2^2 = 0.57$; $P = 0.75$.		$a = 1.09$; $b = 2.24$; $X_2^2 = 0.41$; $P = 0.82$.		$a = 0.99$; $b = 1.89$; $X_2^2 = 0.33$; $P = 0.85$.	

Text 19: Brief Luthers an Katharina Luther vom 25.1.1546.
Text 20: Brief Luthers an Katharina Luther vom 1.2.1546.
Text 21: Brief Luthers an Katharina Luther vom 10.2.1546.

5. Zusammenfassend kann festgestellt werden, daß die Hyperpoisson-Verteilung in ihrer 1-verschobenen Form an alle untersuchten Texte angepaßt werden konnte. Obwohl es sich um Briefe an sehr unterschiedliche Adressaten (Privatpersonen, Fürsten) handelt - hinzu kommt noch die Vorrede zum Kleinen Katechismus als eine ganz andere Textsorte - machen die Luthertexte einen sehr homogenen Eindruck. Dabei ist zu beachten, daß die Luthertexte einem anderen Modell folgen als die Texte des gegenwärtigen Deutschen, die ja der positiven negativen Binomialverteilung entsprechen. Es bleibt nun weiter zu untersuchen, ob die gleichen Befunde für weitere Textsorten Luthers gelten, ob sie womöglich charakteristisch für das Frühneuhochdeutsche insgesamt sind und wie sich die Übergänge zum Verteilungsmodell im gegenwärtigen Deutschen darstellen.

Quellentexte

Text 1 - 11:

Luther, M.: Der Kleine Katechismus für die gemeine Pfarrherr und Prediger. Vorrede. In: *Die Bekenntnisschriften der evangelisch-lutherischen Kirche.* Hg. vom Deutschen Evangelischen Kirchenausschuß. Göttingen [10]1986: Vandenhoeck & Ruprecht, 501 - 507.

Luther, M.: Luthers Briefe. In: *Luthers Werke in Auswahl VI.* Hg. von Hanns Rückert. Berlin [2]1955: de Gruyter.

Text 12 - 21:

Luther, M.: *Luthers Werke. Briefe.* Bd. 7-9, 11. Hg. H. Böhlau und H. Böhlau Nachfolger. Weimar, (ohne Verlag).

Luther, M.: *Ausgewählte Schriften VI.* Hg. v. K. Bornkamm & G. Ebeling. Frankfurt a.M. 1982: Inselverlag.

Literatur

Altmann, G., Erat, E., & Hřebíček, L. (1996). Word Length Distribution in Turkish Texts. In P. Schmidt (Hg.), *Glottometrika 15* (S. 195-204), Trier: WVT.

Ammermann, S. (1997). Untersuchungen zur Wortlängenhäufigkeit in Briefen Kurt Tucholskys. In diesem Band.

Bartels, O., & Strehlow, M., (1997). Zur Häufigkeit von Wortlängen in deutschen Briefen im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts (Bismarck, Brecht, Kafka, Th. Mann, Tucholsky). In diesem Band.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In diesem Band.

Best, K.-H., & Zhu, J., (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Laass, F. (1996). Zur Verteilung von Wortlängen in deutschen Lesebuchtexten. In P. Schmidt (Hg.), *Glottometrika 15* (S. 181-194), Trier: WVT.

Riedemann, H. (1994). *Wortlängen in der Sprache der deutschen und englischen Tages- und Wochenpresse.* Staatsexamensarbeit: Göttingen.

Wolf, H. (1980). *Martin Luther. Eine Einführung in germanistische Luther-Studien.* Stuttgart: Metzler.

Glottometrika 16, 1997, 63-70

Untersuchung zur Wortlängenhäufigkeit in Briefen Kurt Tucholskys

Stefan Ammermann

0. Ein Modell der Wortlängenverteilung im Deutschen

Die vorliegende Untersuchung entstand im Zusammenhang mit dem Projekt „Wortlängenhäufigkeit", das besonders in Bochum und Göttingen betrieben wird. Erkenntnisziel ist die Frage, ob die Häufigkeit, mit der Wörter verschiedener Länge in Texten verwendet werden, mathematisch modelliert werden können oder nicht.

Es gibt einen guten Grund für die Annahme, daß neuere deutsche Texte der positiven negativen Binomialverteilung folgen (Best & Zhu, 1994; Altmann & Best, 1996), deren Formel wie folgt lautet:

$$P_x = \frac{\binom{k+x-1}{x} p^k q^x}{1-p^k}, x = 1,2,3\ldots; k > 0; 0 < p < 1; q = 1-p.$$

In dieser Untersuchung soll nun geprüft werden, ob diese Vermutung sich am Beispiel von Briefen Tucholskys bestätigen läßt. Dabei sind Briefe ein besonders geeignetes Untersuchungsobjekt, da sie in der Regel als sehr homogene Texte aufgefaßt werden können: Sie werden meist aus gegebenem Anlaß niedergeschrieben, ohne daß größere zeitliche Unterbrechungen eintreten und eine spätere Bearbeitung stattfindet. (Zur Untersuchung weiterer Briefe vgl. Bartels & Strehlow in diesem Band).

1. Zur Textbearbeitung

Um eine Analyse von Texten nach der Häufigkeit von Wörtern unterschiedlicher Länge, gemessen in der Zahl der Silben pro Wort, durchführen zu können, müssen die Einheiten „Wort" und „Silbe" bestimmt werden. Aus den verschiedenen Möglichkeiten der Definition von Wort (Lühr, [4]1993:131f) wird das „orthographische Wort" als besonders einfach zu handhaben betrachtet und deshalb zur

Grundlage dieser Arbeit gemacht. Der Bindestrich gilt als Hinweis auf die Einheit des Wortes, das Apostroph wurde zunächst als worttrennend aufgefaßt; der einzige Fall eines dadurch ermittelten nullsilbigen Wortes wurde dann aber bei der Auswertung der Daten vernachlässigt. Man kann mit guten Grund die Auffassung vertreten, daß nullsilbige Einheiten zumindestens phonetisch gesehen zu Bestandteilen ihrer Nachbarwörter werden.

Die Silbe wurde nach dem Vorkommen von Vokalen und Diphthongen bestimmt: d.h. ein „Wort" hat soviele Silben, wie es Vokale oder Diphthonge enthält.

Zur Bearbeitung der Texte ist schließlich noch auf folgende Entscheidung hinzuweisen: Da es sich bei den Texten um Briefe handelt, blieben jeweils nur die Ortsangabe und das Datum unbeachtet. Die Anrede ist als Text zugehörig gewertet, ebenso wie die Unterschrift, die bei Tucholsky oft als ganzer Satz formuliert ist („Hadjö sagt Ihm sein Nungo", vgl. Brief vom 26. Mai 1918). Abkürzungen sind in ihrer laut gelesenen Form gezählt. Also z.B. „Nr." als „Nummer" - 2 Silben. Bei der Auszählung von Zahlwörtern gilt die vereinbarte Regelung, daß sie nach Tausendern, Hundertern und Zehnern in Einzelwörter zerlegt werden. So ist z.B. „25" („fünfundzwanzig") als viersilbig anzusehen und die Jahreszahl „1797" (Text 10) als zwei Wörter („siebzehnhundert siebenundneunzig") mit vier bzw. fünf Silben. Bei der Bewertung von Fällen wie z.B. „Original" (viersilbig) wurde die Bühnenaussprache zugrunde gelegt. Hier werden unsilbische Vokale als silbentragende realisiert (Duden: Aussprachewörterbuch: 53). Weiterhin gilt bei Apostrophierung, wie z.B. bei „in'n" (Text 2): zwei Wörter („in den") - einsilbig und nullsilbig. Anders verhält es sich bei der Zusammenziehung ohne Apostroph (z.B. „ichs" für „ich es"). Dieser Sachverhalt läßt sich ausgesprochen häufig in den vorliegenden Texten finden und wird als jeweils ein Wort in der laut gelesenen Form gewertet.

Von der oben aufgeführten Regelung wurde jeweils dann abgewichen, wenn die Form als vom Autor beabsichtigt zu erkennen war. So etwa bei dem „régime Z.". Hier steht die Abkürzung „Z." für „Zimmermann"[1]. Es kann davon ausgegangen werden, daß Tucholsky diese Abkürzung als scherzhaft karikierend verstanden wissen wollte. Aus diesem Grunde habe ich mich in einigen Fällen für die Beibehaltung solcher Formen entschieden, d.h. „Z" ([tsɛt]) als eine Silbe.

Der geringe französische Textanteil ist gemäß der französischen Aussprache behandelt worden. Die Form „Mély pai-painka" (baltisches Kinderdeutsch = streicheln) in Text 7 wurde wie zwei Wörter behandelt (Mély - 2 Silben, pai-painka - 3 Silben). Zum Schluß sei noch auf die häufige Verwendung von Interjektionen hingewiesen, die ebenfalls in der laut gelesenen Form ausgewertet wurden und keine Probleme bereiteten.

[1] Hierbei handelt es sich um Hauptmann Zimmermann: Kommandeur der Artillerie-Fliegerschule-Ost in Alt-Autz in Kurland, an der Mary Gerold Dienstverpflichtete war.

2. Die computerunterstützte Textanalyse

Die Möglichkeiten der elektronischen Datenverarbeitung (EDV) lassen sich auch im Bereich der Quantitativen Linguistik nutzbar machen.

In folgenden soll das Vorgehen bei der Erhebung der Daten und deren weitere Bearbeitung mit Hilfe des Computers näher beschreiben und erläutern werden.

a. Einscannen der Texte

Um die Texte mit Hilfe des Computers auf ihre Wortlänge hin zu untersuchen, mußten sie zunächst für diesen lesbar gemacht werden. Hierzu boten sich zwei Möglichkeiten an:

1) Eintippen der Texte oder
2) Einscannen der Texte.

Ein Eintippen der Texte hätte einen erheblichen Arbeitsaufwand bedeutet und die Verwendung des Computers in keinem Fall gerechtfertigt.

Das Einlesen des Textes unter Verwendung eines Scanners ist hingegen eine lohnende und zeitsparende Maßnahme, da der Text, ähnlich wie beim herkömmlichen Fotokopieren, abgetastet und direkt in den Computer geleitet wird. Heutige Texterkennungsprogramme sind in der Lage zwischen einer Vielzahl verschiedener Schrifttypen zu unterscheiden und diese beim Lesevorgang mit einer hohen Wahrscheinlichkeit zu identifizieren. Ich selbst habe mit einem Flachbettscanner und der Texterkennung Omnipage™ gearbeitet.

b. Nachbearbeitung

Da kein Texterkennungssystem mit hundertprozentiger Sicherheit arbeitet, ist es notwendig, den Text nach dem Einscannen mit einem Editor oder einem Textverarbeitungsprogramm (z.B. WORD) nachzubearbeiten. Im vorliegenden Fall war ab und zu eine Korrektur der nichterkannten Umlaute vorzunehmen.

Entscheidet man sich nun mit dem Silbenerkennungsprogramm SYLC[2] zu arbeiten, dann gilt als zweiter wichtiger Schritt bei der Nachbearbeitung die Umformung von Gedankenstrichen (hier empfiehlt sich, die Gedankenstriche zu löschen), da SYLC keine Unterscheidung zwischen Bindestrichen und Gedankenstrichen macht. Während der Gedankenstrich Satzteile einschiebt, betont der Bindestrich die Einheit des Wortes. Es handelt sich also um zwei verschiedene Sachverhalte.

[2] Die verwendeten Programme wurden mir von Herrn G. Altmann („Altmann-Fitter"), Herrn M. Diel (SYLC) sowie der Gesellschaft zur wissenschaftlichen Datenverarbeitung (GWDG) zur Verfügung gestellt.

c. Arbeiten mit SYLC und dem „Altmann-Fitter"

Nachdem der Text im ASCII-Format (Standardzeichenformat ohne Formatierung des Textes) abgespeichert war, konnte er mit SYLC bearbeitet werden, indem zunächst die Wortlänge jedes Wortes einmal vom Benutzer eingegeben wird. Jedes wiederholte Auftreten eines Wortes wird automatisch erkannt. Das Programm ist somit „lernfähig", und der Arbeitsaufwand wird mit zunehmendem Wortbestand verringert. Die Daten der Auszählung werden in einer Sammeldatei (SYLC.LOG) gespeichert. Zur weiteren Verarbeitung mußten die Daten jedes einzelnen Textes in einer ihr zugehörigen Datei gespeichert werden.

Die Anpassung der positiven negativen Binomialverteilung wurde iterativ durchgeführt. Die Werte in den Tabellen bedeuten:

- NPx - theoretische Häufigkeit;
- k und p - Parameter der positiven negativen Binomialverteilung;
- X^2 - der Wert des Chiquadrats im Anpassungstest mit der Anzahl der Freiheitsgrade im Index;
- C - Diskrepanzkoeffizent $C = X^2/N$;
- P - die Überschreitungswahrscheinlichkeit.

Die Ergebnisse konnten dann in tabellarischer und graphischer Form einander gegenübergestellt, überprüft und bewertet werden.

3. Zur Textauswahl

Es handelt sich bei den 15 untersuchten Texten um Briefe Kurt Tucholskys an Mary Gerold (seine spätere Frau). Diese entstanden zwischen dem 29. April und dem 30. Juni 1918. Funktionalstilistisch lassen sie sich der Alltagssprache (Konversationsstil) zuordnen. Ihr durchschnittlicher Umfang beträgt 774 Worte (kürzester Text = 353 Worte, längster Text = 1004 Worte). Bei der Textauswahl kam es mir darauf an, Briefe in einer lückenlosen und chronologischen Abfolge auszuzählen, da so dem Problem vorgebeugt wird, daß der Autor seinen Stil im Laufe der Zeit geändert haben könnte. Hieraus hätten sich Differenzen bezüglich der Wortlängenhäufigkeitsverteilung zwischen den Texten ergeben können (vgl. Wimmer, Köhler, Grotjahn & Altmann, 1994). Die Briefe vom 28.5.1918 und vom 5.6.1918 wurden nicht bei der Auszählung berücksichtigt, da es sich hierbei um kurze Prosastücke handelt, bei denen u.a. die persönliche Anrede fehlt.

Auf folgende stilistische Besonderheiten Tucholskys sei noch kurz hingewiesen: Tucholsky paßt seine Orthographie der phonologischen Realisierung an, d.h. er schreibt z.B. „unserm" für „unserem" oder „Lehm" statt „Leben" (beide aus Text 6). Daneben neigt er vereinzelt zu der Bildung von Neologismen wie z.B.

„Sonntagnachmittagskindergesellschaft" (Text 5) oder „Wippwappgang" (Text 6), die das Gesamtbild der Wortlängenverteilung beeinflussen. Zu solchen Ausnahmefällen findet sich jeweils eine Anmerkung unterhalb der entsprechenden Tabelle.

Ein weiteres Merkmal Tucholskys ist die Verwendung französischer Redewendungen und Zitate. Nach meinen Beobachtungen wirkt sich dies aber nicht negativ auf die Modellierbarkeit der Texte aus. Auch hierzu wird ein entsprechender Hinweis in der Tabelle gegeben. Ein Sachverhalt, der für die Untersuchung aber ohne Bedeutung war, ist die eigenwillige Orthographie Tucholskys. Zur Betonung der Aussage vervielfacht er Vokale, wie z.B. bei „puuuhle" (Text 1), und als ironische Anspielung auf die Aussprache Marys vertauscht er den Vokal „e" mit „i" beim Präfix „ge-" (Beispiel: „gidacht" in Text 9).

4. Die Datenerhebung hat folgende Ergebnisse erbracht:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	492	494.48	592	594.96	208	208.77
2	239	230.14	298	279.00	101	94.56
3	76	84.30	73	95.39	24	29.18
4	28	26.70	34	26.56	5	6.90
5	8	7.66	5	6.38	3	1.59
6	3	2.73	1	1.37	-	-
7	-	-	1	0.33	-	-
	$k = 4.5419$; p 0.8320; $X^2_3 = 1.28$; $P = 0.73$.		$k = 9.6735$; $p = 0.9121$; $X^2_3 = 9.00$; $P = 0.029$; $C = 0.009$.		$k = 44.2949$; $p = 0.9800$; $X^2_2 = 3.14$; $P = 0.21$.	

Text 1: Brief vom 29.4.1918.
Text 2: Brief vom 6.5.1918. Anmerk.: „in'n" gewertet als zwei Wörter in- und nullsilbig). Bei der Berechnung von NP_x wurde das nullsilbige Wort nicht berücksichtigt und deshalb in der Tabelle nicht aufgeführt.
Text 3: Brief vom 7.5.1918.

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	211	211.61	527	529.96	492	457.89
2	96	93.37	273	260.45	212	214.19
3	30	33.34	87	101.11	76	83.59
4	12	10.51	38	34.04	39	29.38
5	4	4.17	11	10.40	7	9.64
6	-	-	3	2.97	3	4.30
7	-	-	0	0.80	-	-
8	-	-	0	0.21	-	-
9	-	-	0	0.05	-	-
10	-	-	1	0.02	-	-
	$k = 3.6708$; $p = 0.8111$; $X_2^2 = 0.63$; $P = 0.73$.		$k = 4.4066$; $p = 0.8182$; $X_4^2 = 3.09$; $P = 0.54$.		$k = 2.9778$; $p = 0.7648$; $X_3^2 = 5.01$; $P = 0.17$.	

Text 4: Brief vom 17.5.1918. Anmerk.: Abkürzung "Z." ("Zimmermann") bleibt beibehalten.
Text 5: Brief vom 20.5.1918. Anmerk.: Sonntagnachmittagskinder-gesellschaft".
Text 6: Brief vom 22.5.1918. Anmerk.: Text in französischer Sprache (55 Wörter).

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	527	528.35	503	506.65	571	569.97
2	265	256.19	251	241.92	246	242.94
3	78	89.95	87	88.65	67	74.24
4	31	25.25	23	27.57	24	18.21
5	2	6.24	7	7.65	2	4.65
6	2	1.35	3	1.95	-	-
7	3	0.36	1	0.6	-	-
	$k = 10.6080$; $p = 0.9165$; $X_2^2 = 3.16$; $P = 0.21$.		$k = 5.6141$; $p = 0.8556$; $X_3^2 = 2.02$; $P = 0.57$.		$k = 12.2621$; $p = 0.9357$; $X_2^2 = 4.10$; $P = 0.13$.	

Text 7: Brief vom 24.5.1918. Anmerk.: Mély pai-painka (baltisches Kinderdeutsch).
Text 8: Brief vom 26.5.1918.
Text 9: Brief vom 28.5.1918.

	Text 10		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	481	483.08	522	521.23	558	558.49
2	235	224.72	250	249.69	295	291.67
3	73	85.51	87	91.01	101	103.36
4	33	28.92	32	27.96	31	27.96
5	10	9.05	7	7.63	2	6.15
6	2	2.68	2	2.48	2	1.37
7	0	0.76	-	-	-	-
8	1	0.28	-	-	-	-
	$k = 3.4039$; $p = 0.7887$; $X_4^2 = 3.16$; $P = 0.53$.		$k = 6.0770$; $p = 0.8646$; $X_3^2 = 0.91$; $P = 0.82$.		$k = 54.8284$; $p = 0.9813$; $X_3^2 = 3.52$; $P = 0.32$.	

Text 10: Brief vom 30.5.1918.
Text 11: Brief vom 3.6.1918.
Text 12: Brief vom 10.6.1918.

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	557	561.15	542	544.89	531	539.67
2	295	279.69	288	266.14	281	252.48
3	80	95.02	71	96.78	72	89.29
4	28	24.74	38	29.15	21	26.48
5	6	5.26	5	7.69	8	6.95
6	1	1.14	3	2.35	3	1.66
7	-	-	-	-	1	0.47
	$k = 43.7058$; $p = 0.9777$; $X_3^2 = 3.79$; $P = 0.29$.		$k = 7.5669$; $p = 0.8860$; $X_3^2 = 12.48$; $P = 0.01$; $C = 0.0132$.		$k = 6.4670$; $p = 0.8747$; $X_3^2 = 9.64$; $P = 0.02$; $C = 0.0105$.	

Text 13: Brief vom 14.6.1918.
Text 14: Brief vom 24.6.1918.
Text 15: Brief vom 30.6.1918.

5. Ergebnisse und Perspektiven

Die Bewertung der Ergebnisse stützt sich hauptsächlich auf P, den Wert, der angibt, mit welcher Wahrscheinlichkeit der errechnete X^2-Wert erreicht oder überschritten wird. Er gilt dann als zufriedenstellend, wenn $P \geq 0{,}05$. Dies ist in einigen Fällen (Text 2, 14, 15) nicht der Fall. Hier kann ersatzweise auf den Diskrepanzkoeffizienten $C = X^2/N$ zurückgegriffen werden. Die Anpassung gilt dann als zufriedenstellend, wenn $C \leq 0{,}02$; dies ist auch in den oben genannten 3 Briefen der Fall. Damit läßt sich feststellen, daß die untersuchten Briefe Tucholskys die Annahme bestätigen, daß die positive negative Binomialverteilung ein valides Modell für eine Vielzahl deutscher Texte darstellt.

Abschließend sei noch auf die Möglichkeit verwiesen, Tucholskys Gedichte und Kurzprosa auf eine eventuelle Übereinstimmung der Wortlängenverteilung mit den hier vorliegenden Ergebnissen hin zu untersuchen. Dies könnte Gegenstand einer zukünftigen Arbeit sein und darüber hinaus helfen, die Datenbasis des Projekts zu vergrößern. Bei diesem Vorhaben würde sich der Vorteil ergeben, daß dem Silbenerkennungsprogramm SYLC der „Wortschatz" Tucholskys bereits teilweise bekannt ist und dadurch eine schnellere Datenaufnahme möglich ist.

Literatur

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In P. Schmidt (Hg.), *Glottometrika 15* (S. 166-180), Trier: WVT.

Bartels, O., & Strehlow, M. (1997). Zur Häufigkeit von Wortlängen in deutschen Briefen im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts (Bismarck, Brecht, Kafka, T. Mann, Tuchlosky). In diesem Band.

Best, K.-H. (1997) Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In diesem Band.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Diel, M. (1994). Wortlängen in deutschen und dänischen Texten. *Manuskript.*

Duden. Aussprachewörterbuch (^{3}1990). Mannheim-Wien-Zürich: Dudenverlag.

Lühr, R. (^{4}1993). *Neuhochdeutsch.* München: Fink.

Tucholsky, K. *Unser ungelebtes Leben - Briefe an Mary.* Herausgegeben von Fritz J. Raddatz. Reinbek bei Hamburg: Rowohlt 1990.

Wimmer, G. , Köhler, R. , Grotjahn, R. , & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of quantitative linguistics,* 1, 98-106.

Glottometrika 16, 1997, 71-76

Zur Häufigkeit von Wortlängen in deutschen Briefen im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts

(Bismarck, Brecht, Kafka, T. Mann, Tucholsky)

Olaf Bartels, Michael Strehlow

0. Im Rahmen des Projektes zur Untersuchung der Häufigkeit von Wortlängen verschiedener Sprachen an der Universität Göttingen 1993/94 wurden u.a. auch deutschsprachige Briefe bekannter Persönlichkeiten auf ihre Wortlängenhäufigkeit hin untersucht. Briefe erschienen uns deshalb als geeignete Textsorte, weil sie meist spontan und ohne Unterbrechung entstehen und in Umfang und thematischer Geschlossenheit dem Forschungsprojekt gegenüber angemessen sind. Es wird fast ausschließlich der Funktionalstil der Alltagssprache verwendet; Themen spezieller Art (z.B. Literatur, Politik) werden nicht mit ausschließlich fachlichen Termini abgehandelt, sondern auf alltagssprachlichem Niveau gehalten.

Die ausgesuchten Texte stammen a) aus der Mitte des 19. Jh. (Bismarck) und b) aus der ersten Hälfte des 20. Jh. (Kafka, Tucholsky, T. Mann, Brecht). Es wurden Briefe berücksichtigt, deren Wortzahl deutlich unter der kritischen Grenze von ca. 2000 Wörtern (Best & Zhu, 1994:20) lag. Für jeden Text wurde festgestellt, wieviele Wörter mit einer, zwei, drei etc. Silben er enthält. Zur Bestimmung von „Wort" wählten wir das „orthographische Wort", „Silbe" wurde nach der Zahl der im Wort vorkommenden Vokale bestimmt; wir folgten damit den gleichen Prinzipien wie schon Best & Zhu (1994:20).

1. Obwohl unsere Texte aus verschiedenen Zeitabschnitten stammen, prüfen wir, ob sie wie andere neuere deutsche Texte der nullgestutzten (positiven) negativen Binominalverteilung folgen, deren Formel wie folgt lautet:

$$P_x = \frac{\binom{k+x-1}{x} p^k q^x}{1-p^k}, \quad x = 1, 2, \ldots; \quad k > 0; \quad 0 < p < 1; \quad q = 1 - p.$$

Die Anpassung wird als zufriedenstellend betrachtet, wenn $P(X^2) \geq 0{,}05$ bzw. $C \leq 0{,}02$ ist. Die Berechnungen wurden mit dem Altmann-Fitter in Bochum durchgeführt.

Auf Besonderheiten der Wort- und Silbenbestimmung wird bei der Vorstellung der einzelnen Texte hingewiesen. Die Bismarckbriefe wurden vollständig, einschließlich Datum, Ortsangabe, Anrede und Gruß bearbeitet; bei den neueren Briefen wurde auf Ortsangabe und Datum verzichtet, da diese Angaben von den Autoren sehr unterschiedlich gehandhabt wurden. Diese etwas uneinheitliche Art der Datenerhebung wurde nicht nachträglich verändert, da sie sich bei den Rechnungen nicht störend bemerkbar gemacht hat.

2. Die Ergebnisse stellen sich wie folgt dar:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	610	614.72	341	339.97	492	494.28
2	324	310.78	225	225.83	321	310.82
3	96	106.50	109	111.62	125	137.72
4	29	27.82	49	45.68	53	48.23
5	7	5.91	18	16.36	15	14.20
6	1	1.27	3	5.30	3	3.65
7	-	-	0	1.59	0	0.84
8	-	-	2	0.65	1	0.26
	$k = 58.6673$; $p = 0.9831$; $X_3^2 = 1.94$; $P = 0.59$.		$k = 7.6098$; $p = 0.8457$; $X_4^2 = 1.49$; $P = 0.83$.		$k = 16.5504$; $p = 0.9283$; $X_4^2 = 2.16$; $P = 0.71$.	

Text 1: Bismarck, Brief Nr. 36: An die Braut, 15.5.1847 (Hans Rothfels, Hg., 1955: 99ff)
Anmerkung: „B." (für „Bismarck") einsilbig gewertet.

Text 2: Bismarck, Brief Nr.43: An die Redaktion der Magdeburger Zeitung, 20.4.1848 (Textquelle wie Text 1, S. 110ff).
Anmerkung: Die achtsilbigen Wörter sind Jahreszahlen. Sie wurden trotz einiger Bedenken als unzerlegte Wörter beibehalten, da es sich nur um wenige Fälle handelt.

Text 3: Bismarck, Brief Nr. 102: An Leopold von Gerlach, 25.10.1854 (Textquelle wie Text 1, S. 193ff).

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	489	488.30	181	183.80	145	147.12
2	284	283.36	92	79.56	101	96.38
3	116	120.77	27	35.37	40	42.46
4	47	42.17	12	15.93	14	14.15
5	12	12.77	12	7.23	4	3.80
6	3	3.47	4	6.11	1	1.09
7	0	0.86	-	-	-	-
8	1	0.30	-	-	-	-
	$k = 8.8343$; $p = 0.8820$; $X_4^2 = 0.87$; $P = 0.93$.		$k = 0.8497$; $p = 0.5320$; $X_2^2 = 5.47$; $P = 0.06$.		$k = 111.6594$; $p = 0.9884$; $X_3^2 = 0.41$; $P = 0.94$.	

Text 4: Bismarck, Brief Nr. 148: An Albrecht von Roon, 2.7.1861 (Textquelle wie Text 1, S. 276ff). Anmerkung: „Sr." (= „Seiner") wurde zweisilbig gewertet; „wenn's": „s" wird wie andere nullsilbige Wörter als phonetischer Bestandteil des Nachbarwortes betrachtet und nicht als eigene Längenklasse aufgeführt.

Text 5: Kafka, Brief an Milena, November 1920, Prag (aus: Kafka 1951) Anmerkung: „A", „B", „C" wurden als einsilbige Wörter gewertet.

Text 6: Kafka, Brief an Ottla, 4.-5.9.1917, Prag (Textquelle wie Text 5).

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	174	174.17	226	224.91	265	264.94
2	82	80.62	115	114.13	116	112.54
3	30	31.86	42	45.48	34	40.12
4	12	11.51	20	15.65	16	12.93
5	5	3.93	2	4.87	4	3.94
6	1	1.91	1	1.41	1	1.57
7	-	-	1	0.55	-	-
	$k = 2.5599$; $p = 0.7399$; $X_3^2 = 0.86$; $P = 0.84$.		$k = 4.6182$; $p = 0.8193$; $X_3^2 = 3.18$; $P = 0.36$,		$k = 2.8647$; $p = 0.7802$; $X_3^2 = 1.97$; $P = 0.58$.	

Text 7: Kafka, Brief an Milena, 8.7.1920, Prag (Textquelle wie Text 5).

Text 8: Tucholsky, Brief an Bernays, 30.9.1929 (aus: Tucholsky 1977).

Text 9: Tucholsky, Brief an Matzlein, 17.10.1918. (Textquelle wie Text 8).

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	137	131.54	183	184.67	170	171.44
2	64	73.92	101	94.59	101	93.60
3	37	34.14	29	35.33	31	38.39
4	19	14.06	12	10.74	16	13.14
5	4	5.36	4	3.67	2	3.96
6	1	2.98	-	-	2	1.47
	$k = 3.2968$; $p = 0.7384$; $X_3^2 = 5.17$;. $P = 0.16$		$k = 9.6615$; $p = 0.9039$; $X_2^2 = 1.76$; $P = 0.41$.		$k = 6.8723$; $p = 0.8613$; $X_3^2 = 3.82$; $P = 0.28$.	

Text 10: Tucholsky, Brief an Julius Bab, 1.4.1927 (Textquelle wie Text 8).
Text 11: Tucholsky, Brief an Herrn Vordtriede, 14.9.1930 (Textquelle wie Text 8).
Text 12: Th. Mann, Brief an Hermann Hesse, 12.5.1935, Küsnacht (aus: Mann 1968).

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	144	146.16	144	146.00	316	319.65
2	78	69.96	75	66.72	212	200.73
3	20	25.38	21	26.80	89	98.06
4	6	7.73	8	10.02	43	41.07
5	3	2.09	7	5.46	14	15.48
6	0	0.51	-	-	6	5.40
7	0	0.11	-	-	1	1.77
8	1	0.06	-	-	0	0.55
9	-	-	-	-	2	0.29
	$k = 6.3012$: $p = 0.8689$; $X_2^2 = 3.05$; $P = 0.22$.		$k = 2.1425$; $p = 0.7091$; $X_2^2 = 3.16$; $P = 0.21$.		$k = 4.9907$; $p = 0.7904$; $X_4^2 = 1.88$; $P = 0.76$.	

Text 13: T. Mann, Brief an Hermann Hesse, 10.8.1947, Zürich (Textquelle wie Text 12). Anmerkung: Zahlen: „28., 25., 75." fünfsilbig gewertet.
Text 14: T. Mann, Brief an Hermann Hesse, 1.7.1949, Graubünden, Hotel Schweizerhof (Textquelle wie Text 12).
Text 15: T. Mann, Brief an Hermann Hesse, 27.11.1931, München (Textquelle wie Text 12).

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	414	419.32	390	390.06	445	441.55
2	231	218.17	183	180.20	244	243.71
3	93	95.06	64	69.77	108	120.24
4	29	37.40	29	24.40	67	55.80
5	17	13.76	7	7.99	31	24.92
6	4	4.83	2	2.49	8	10.83
7	2	1.63	0	0.75	1	4.62
8	0	0.54	1	0.34	0	1.94
9	0	0.17	-	-	0	0.80
10	1	0.12	-	-	1	0.59
	$k = 2.9029$: $p = 0.7334$: $X_4^2 = 3.79$; $P = 0.44$.		$k = 2.8885$; $p = 0.7624$; $X_4^2 = 1.61$; $P = 0.81$,		$k = 1.9336$; $p = 0.6237$; $X_6^2 = 10.62$; $P = 0.10$.	

Text 16: Th. Mann, Brief an Hermann Hesse, Pacific Palisades, California, 13.7.1941 (Textquelle wie Text 12). Anmerkung: 20 Wörter aus dem Englischen, die aber getrennt gezählt der Normalverteilung entsprechen; daher der Rechnung integriert.
Text 17: Brecht, Brief an Max Hohenester (aus: Brecht 1981). Anmerkung: „u[nd]" wird als „und" gelesen. Selbst wenn es bei Brecht nur als „u" gelesen worden wäre, müßte dafür eine Silbe gerechnet werden. Es ändert sich also nichts an der Zählung.
Text 18: Brecht, Brief an Karl Korsch (Textquelle wie Text 17).

Zusammenfassend kann festgestellt werden: Alle Texte lassen sich mit der nullgestutzten (positiven) negativen Binominalverteilung modellieren, obwohl sie von verschiedenen Autoren stammen und obwohl Bismarcks Briefe deutlich früher geschrieben wurden als die anderen. Es zeigt sich damit wieder einmal, daß das Deutsche relativ homogen wirkt, selbst wenn man Texte einer längeren Zeitspanne wählt.

4. Literatur:

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Brecht, B. (1981). *Briefe*. Hg. v. Glaeser, G., Frankfurt am Main: Suhrkamp.

Duden. Die deutsche Rechtschreibung (1991). Mannheim, Wien, Zürich: Dudenverlag.

Kafka, F. (1951). *Gesammelte Werke. Bd. 5. Briefe an Milena*. Hg. v. Brod, M., Frankfurt am Main: Fischer.

Mann, T. (1968). *Hermann Hesse und Thomas Mann - Briefwechsel*. Frankfurt am Main: Suhrkamp.

Rothfels, H. (1955). *Bismarck-Briefe*. Göttingen: Vandenhoeck & Ruprecht.

Tucholsky, K. (1977). *Briefe aus dem Schweigen 1932-1935*. Hg. v. Tucholsky, M.G. & Huonker, G., Reinbek: Rowohlt.

Glottometrika 16, 1997, 77-87

Die Wortlängenhäufigkeiten von deutschsprachigen naturwissenschaftlichen Publikationen

Gabi Behrmann

1. Texte zeigen unterschiedliche Häufigkeiten, mit denen Wörter verschiedener Länge auftreten. Diese Untersuchung geht davon aus, daß die Wortlängen nach bestimmten Gesetzmäßigkeiten verteilt sind. Zur Modellierung von deutschsprachigen Kurzprosatexten hat sich die 0-gestutzte oder positive negative Binomialverteilung bereits bewährt (Best & Zhu, 1994). Gleiches gilt für deutschsprachige Pressetexte (Best, 1997). Untersuchungen im Chinesischen (Best & Zhu, 1994) und Schwedischen (Best, 1996) zeigen jedoch Abweichungen von dieser Verteilung. So folgen beispielsweise die untersuchten schwedischen Pressetexte der Poisson-Verteilung. In verschiedenen Sprachen konnten also bereits unterschiedliche Verteilungen nachgewiesen werden. Man muß mit der Möglichkeit rechnen, daß auch unterschiedliche Textsorten derselben Sprache verschiedenen Modellen folgen können, so daß die Textsorten bzw. Funktionalstile einer Sprachgemeinschaft daraufhin zu untersuchen sind, welche der vielen möglichen Verteilungen geeignet sind, sie zu modellieren. Dies soll hier anhand der Darstellung von deutschsprachigen naturwissenschaftlichen Publikationen unternommen werden.

2. Die Texte für die Datenerhebung sind den Jahresbänden von *Natur und Museum* (1929-1933: Bd. 59-63) und *Natur und Volk* (1934-1936: Bd. 64-66, 1938-1941: Bd. 68-71, 1949-1952: Bd. 79-82) entnommen. Die eher populärwissenschaftlichen Veröffentlichungen der Senckenbergischen Naturforschenden Gesellschaft in Frankfurt/M. erscheinen im Selbstverlag seit 1850 im *Bericht* und seit 1922 bis heute in *Natur und Museum*. Von 1934 bis 1961 hieß die Zeitschrift vorübergehend *Natur und Volk*.

Es handelt sich aber um ein und dieselbe Reihe, die zwischen der naturwissenschaftlichen Forschung und der Bevölkerung zu vermitteln sucht, so daß sowohl Fachleute als auch Laien als potentielle Leser in Betracht kommen. Dementsprechend weisen die fachlichen Beiträge in diesen Heften, die monatlich bzw. alle zwei Monate erscheinen, trotz der fachsprachlichen Eigenschaften einen nur gemäßigt wissenschaftlichen Stil auf: So werden Fachtermini durch Er-

läuterungen allgemein verständlich gemacht oder nur in reduziertem Maße verwendet. Es wird ein breites, nicht fachspezifisches Publikum, aber ein durchaus fachlich interessierter Leser vorausgesetzt. Zudem dienen die Publikationen der interdisziplinären Kommunikation innerhalb der jeweiligen Fachwissenschaften. Die einzelnen Beiträge entsprechen dem aktuellen Forschungsstand eines Fachgebietes und haben dann einen sehr speziellen Charakter, oder sie setzen sich mit allgemeineren Themen auseinander.

Eine Texthomogenität ist wegen der starken fachsprachlichen Einflüsse unterschiedlicher Wissenschaftsbereiche nicht zu erwarten.

Es wurden Texte der Zoologie, Botanik, Geographie, Geologie, Anthropologie, Vor- und Frühgeschichte, Völkerkunde, Chemie und Physik untersucht. Damit ist nahezu das gesamte Spektrum der in der Zeitschrift vertretenen Wissenschaften abgedeckt. Die meisten ausgewählten Texte sind jedoch zoologische Publikationen, da hier lateinische Termini, Zahlen, Symbole und Formeln vergleichsweise weniger gehäuft auftreten.

Der Umfang der bearbeiteten Artikel beträgt um die 1000 Wörter. Das entspricht einer nicht zu hohen Wortanzahl unter Beachtung der wachsenden Textinhomogenität bei mehr als 2000 Wörtern (Best & Zhu, 1994:20) sowie gleichzeitig eines genügend großen Umfanges, um die vorkommenden lateinischen Termini und Zahlen möglichst wenig ins Gewicht fallen zu lassen. Die Texte sind alle deutschsprachig. Obwohl jeweils einzelne Wissenschaftler ihre Publikationen verfaßt haben, sind wegen möglicher Korrekturen durch die Schriftleitung jeweils mehrere Autoren anzunehmen.

Unter Wahrung der genannten Kriterien wurden die Texte willkürlich ausgewählt.

3. Die Auszählung der Texte erfolgte durch eine Bearbeiterin nach immer derselben Methode. Die Texte wurden in mehreren separaten Durchgängen abschnittsweise bearbeitet. Hat eine erste Kontrollzählung andere Werte als die vorangehende Zählung ergeben, wurde eine erneute Zählung durchgeführt. Schließlich wurden alle Werte für einen Text aufsummiert.

Bei der Zählung wurde nur der laufende Text berücksichtigt, also ohne Überschriften (darunter fallen: Titel, Untertitel, Autorennennungen und Ergänzungen wie z.B. Zahl der Abbildungen), nachstehende Literaturhinweise, Autorenunterschriften, Fußnotenzeichen innerhalb des Textes sowie Fußnotentext und Abbildungsunterschriften. Die Zählung beginnt also mit dem im Druck deutlich abgesetzten Textblock und endet mit demselben. Innerhalb des fortlaufenden Textes wurde in Klammern oder innerhalb von Gedankenstrichen stehende Ergänzungen und Abbildungshinweise mitberücksichtigt, da man diese im Text üblicherweise fortlaufend mitliest. Gemäß der Orientierung an der gelesenen Realisation des Textes blieben deshalb Fußnotenzeichen unberücksichtigt, weil sie gewöhnlich auch nicht mitgelesen, sondern im Lesefluß ausgespart werden.

Das „Wort" wurde nach dem „graphematischen Wort" definiert: „(...) als wahrnehmbare Einheit des geschriebenen Textes (...). Man erkennt [es] in einem Text an Zwischenräumen bzw. an einigen Sonderzeichen (...)" (Bünting & Bergenholtz, 1989:39).

Die Wortlänge wurde nach der Zahl der Silben bestimmt. Die Zahl der Silben entspricht dabei der Zahl der Vokale bzw. Diphthonge in einem Wort. Um eine einfache Operationalisierung bei der Silbenzahlbestimmung zu ermöglichen, wurden die Silben buchstabengetreu bestimmt: Demnach „Studium" [stu:dium] als dreisilbig (Stu-di-um), was der Bühnenaussprache (DUDEN, 1974) entspricht, bei der unsilbische Vokale als silbentragende realisiert werden. Für Wörter lateinischer oder griechischer Herkunft gilt dasselbe, also z.B. „Geologie" als viersilbig (Ge-o-lo-gie) oder „Element" als dreisilbig (E-le-ment). Auch bei afrikanischen Ortsnamen erfolgte die Silbenzahlbestimmung nach der Zahl der Vokale, also „Ufiome" als viersilbig (U-fi-o-me) und „Ngoron" als zweisilbig (Ngo-ron). Das entspricht der deutschsprachigen Realisation einzelner Vokale und der sog. Definition einer Silbe. Die abweichende Silbenstruktur der Herkunftsprache von Fremdwörtern bleibt bei der Untersuchung von deutschen Texten unberücksichtigt. Doppelvokale wurden entsprechend ihrer gelesenen Form gewertet, demnach „Saal" [z a: l] als einsilbig, aber „Zoologie" [tso'ologi:] als viersilbig (Zo-o-lo-gie).

Zahlwörter wurden entsprechend der orthographischen Konvention als ein Wort gezählt, wobei die gelesene Form berücksichtigt wurde. Das ist für Jahreszahlen und Ziffern unproblematisch, entsprechend gilt für „4½" als „vier-einhalb": 1 Wort, „100 m" als Maß „ein-hun-dert Me-ter": 2 Wörter, „50facher" zu sammengeschrieben als „fünf-zig-fach-er": 1 Wort, „1913/1915" als Jahreszahlenkombination „neun-zehn-hun-dert-drei-zehn neun-zehn-hun-dert-fünf-zehn": 2 Wörter. Der Schrägstrich wurde in der gelesenen Form nicht realisiert.

Das entspricht der grundsätzlichen Behandlung von Interpunktions- und Sonderzeichen im Satz, die beim Lesen sprachlich nicht realisiert werden. Eine Ausnahme stellen Interpunktions- und Sonderzeichen innerhalb von Zahlen dar, wie etwa: „2,50" als „zwei Kom-ma fünf-zig": 3 Wörter, „(der) 8. (Hinterleibs-Ring)" als „ach-te": 1 Wort, „1-3" als „eins bis drei": 3 Wörter, „1:25 000" als Maßstab „eins zu fünf-und-zwan-zig-tau-send": 3 Wörter. In diesen Fällen werden sie durchaus sprachlich realisiert und müssen somit auch mitgezählt werden.

Abkürzungen werden entsprechend der gelesenen Form berücksichtigt, das entspricht meistens der ausgeschriebenen Form. Also: „usw." als „und so weiter": 3 Wörter, „d.h." als „das heißt": 2 Wörter, „Abb." als „Ab-bil-dung": 1 Wort. Die Abkürzungen von Flüssen hinter Städtenamen sowie von Titeln oder Anredeformen vor Namen wurden ihrer gelesenen Form nach, also ebenfalls wie ausgeschrieben, so behandelt: „(Königsberg)/Pr." als „Pre-gel": 1 Wort. Der Schrägstrich wird hierbei nicht sprachlich artikuliert und dementsprechend auch nicht mitgezählt. Ferner „Prof." als „Pro-fes-sor": 1 Wort, "Mr" als "Mis-ter": 1

Wort. Anders wurden Kürzel von Vornamen und Nachnamenbestandteilen behandelt, da diese in der Regel auch nur als Buchstabenkürzel gelesen werden, also „E. (Menner)“ als „E“: 1 Wort, „v. (Frisch)“ als „von“: 1 Wort. Aber in ausgeschriebener Form wird realisiert: “Mc (Cown)” als “Mac”: 1 Wort.

Namenkürzel der Benenner einer Pflanze oder eines Tieres schließen sich lt. Konvention in der Systematik dem lateinischen Gattungs- und Artennamen an. Damit ist innerhalb der internationalen Kommunikation sichergestellt, welcher Autor bei der Artennennung zitiert wird, was für eine eindeutige Identifikation der besprochenen Art notwendig ist. Die Kürzel werden gewöhnlich als Buchstabenkürzel gelesen, demnach: „L.“ als 1 Wort (steht für „Linné“), „HTG“ als 3 Wörter (steht für eine Kombination von drei Namen), „O.S.“ als 2 Wörter (steht für eine Kombination von zwei Namen), „FALL“ als 1 Wort (steht für einen Namen, lies: [f a l]).

Bereits als ein graphematisches Wort gekennzeichnet sind Fälle wie „beim“ = „bei dem“ und „ins“ = „in das“.

Bindestrichwörter wurden als ein Wort gezählt. Der Bindestrich wird damit seiner Funktion nach dem Trennungsstrich gleichgesetzt, daß er nämlich die graphematische Einheit eines Wortes anzeigt. Die Komposita „Tier-Riesen“ oder „Hinterleibs-Ring“ also als 1 Wort. Das gleiche gilt für Namenkomposita mit Bindestrich, wie z.B. „Pangani-Fälle“. Anders wurden Namenkomposita aus mehreren graphematischen Wörtern, die nicht mit einem Bindestrich verbunden sind, wegen des Kriteriums des graphematischen Wortes als mehrwortiges Syntagma behandelt. Als 2 Wörter zählen so z.B. „Wiesbadener Naturhistorisches Museum“.

Lateinische Gattungs- und Artennamen treten immer gemeinsam auf, wie beispielsweise in „Lachnus pichtae“, wobei lt. Konvention die Gattungsbezeichnung immer der Artenbestimmung vorangestellt ist. Sie wurden wegen ihrer graphematischen Kennzeichnung als 2 Wörter gewertet.

Aus methodischer Konsequenz wurden Schreibformen wie „Ngoron goro-Krater“ als 2 Wörter („Ngoron“ und „goro-Krater“) gezählt.

Der Bindestrich kennzeichnet in Fällen wie „Pflanzen- und Tierwelt“ die Zusammengehörigkeit von „Pflanzen-“ mit dem zweiten Kernmorphem „-welt“, das hier zur Vereinfachung (da ökonomischer im Schreib- und Lesefluß) weg- gelassen wurde. Dieses entspricht der üblichen Schreibweise von Ausdrücken aus mehreren kombinierten Determinativkomposita, bei denen das eine Kernmorphem wiederholt auftritt. Da die Leseform der geschriebenen Vorlage entspricht (denn man liest tatsächlich nicht „Pflanzenwelt und Tierwelt“, sondern spart „-welt“ beim ersten Kompositum aus, gleichzeitig wird auch der Bindestrich sprachlich nicht realisiert), wurde „Pflanzen- und Tierwelt“ als aus 3 Wörtern bestehend aufgenommen. Entsprechend wurde „Luftschwere und -feuchtigkeit“, wo das gleiche Phänomen in umgekehrter Form auftritt, als Syntagma mit 3 Wörtern gewertet.

Orthographische Varianten wie z.B. „sodaß“ oder „so daß“ wurden in der vom Autoren verwendeten Form aufgenommen, also als 1 Wort im ersten, als 2 Wörter im zweiten Fall. Dasselbe trifft für auftretende Orthographie- oder Druckfehler zu: Die Zählung erfolgte immer entsprechend der schriftlich realisierten Form.

4. Für die erarbeiteten Daten wurde nach einer Verteilung gesucht, die sich zur Modellierung der Texte eignet. Für deutschsprachige naturwissenschaftliche Publikationen hat sich die positive Singh-Poisson-Verteilung bewährt:

$$P_x = \begin{cases} 1-\alpha+\dfrac{\alpha\, a\, e^{-a}}{1-e^{-a}}, & x=1 \\ \dfrac{\alpha\, a^x\, e^{-a}}{x!\,(1-e^{-a})}, & x=2,3,4,\ldots \end{cases}$$

für $0 < a$ und $0 < \alpha < 1$.

Die Anpassung gilt als zufriedenstellend, wenn $P \geq 0.05$; sie ist noch akzeptabel bei $0.01 \leq P \leq 0.05$; bei längeren Texten ist auch das Diskrepanzmaß $C \leq 0.02$ ein geeignetes Prüfkriterium.

5. Ergebnisse der Datenerhebung:

Erläuterungen zu den Tabellen:

x - Anzahl der Silben pro Wort
n_x - Anzahl der Wörter mit x Silben (absolute Häufigkeiten): beobachtet
NP_x - Anzahl der Wörter mit x Silben: berechnet
a, α - Parameter
X^2 - Werte des Chiquadrates
FG - Freiheitsgrade
P - Überschreitungswahrscheinlichkeit für das entsprechende Chiquadrat
C - Abweichungskoeffizient: $C = \frac{X^2}{N}$

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	502	496.63	400	395.68	462	461.76
2	273	262.37	216	218.72	231	238.02
3	163	171.74	135	144.25	161	150.04
4	71	84.31	78	71.35	80	70.93
5	30	33.11	36	28.23	15	26.83
6	15	10.84	5	9.31	7	8.46
7	5	3.04	0	2.63	3	2.99
8	3	0.75	1	0.86	-	-
9	1	0.23	-	-	-	-
	$a = 1.9637$; $\alpha = 0.7842$; $X^2_3 = 9.005$; $P = 0.0292$ $C = 0.0085$.		$a = 1.9786$; $\alpha = 0.7996$; $X^2_4 = 7.174$; $P = 0.1270$.		$a = 1.8911$; $\alpha = 0.7810$; $X^2_4 = 7.633$; $P = 0.1060$.	

Text 1: E. Scharrer, Bewegungsvorgänge in der Netzhaut des Wirbeltierauges beim Sehen. *Natur und Museum* 59/1929: 99-103.
Text 2: M. Planck, Aus der neuen Physik. *Natur und Museum* 59/1929: 105-107.
Text 3: P. Hirsch, Der Teepilz. *Natur und Museum* 60/1930: 234-236.

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	424	425.73	370	368.20	488	487.36
2	214	216.29	214	216.18	283	273.02
3	172	151.51	158	161.00	160	177.10
4	66	79.60	103	89.93	105	86.16
5	30	33.45	30	40.19	18	33.53
6	11	11.72	18	14.97	10	10.88
7	4	3.52	2	4.78	5	3.02
8	2	1.24	1	1.33	2	0.74
9	-	-	1	0.47	0	0.16
10	-	-	-	-	1	0.09
	$a = 2.1015$; $\alpha = 0.7618$; $X^2_5 = 6.133$; $P = 0.2934$.		$a = 2.2343$; $\alpha = 0.8052$; $X^2_5 = 6.831$; $P = 0.2335$.		$a = 1.9460$; $\alpha = 0.8071$; $X^2_1 = 2.328$; $P = 0.1270$.	

Text 4: K. Hummel, Südafrikanische Landformen. *Natur und Museum* 60/1930: 537-545.
Text 5: M. Galladé, Die Geologie im Wiesbadener Museum. *Natur und Museum* 61/1931: 321-324.
Text 6: H. Schmitz, Über den Wuchsstoff der Pflanzen. *Natur und Museum* 62/1932: 56-59.

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	543	552.48	461	458.95	375	372.10
2	295	297.48	292	287.77	224	222.94
3	165	190.01	171	189.68	136	146.49
4	84	91.03	112	93.77	74	72.19
5	58	34.89	39	37.08	35	28.46
6	25	11.14	8	12.22	9	9.35
7	10	3.05	1	4.57	1	2.63
8	1	0.95	-	-	1	0.84
	$a = 1.9162$; $\alpha = 0.7951$; $X^2_1 = 13.169$; $P = 0.0003$; $C = 0.0112$.		$a = 1.9774$; $\alpha = 0.8451$; $X^2_4 = 9.772$; $P = 0.0445$; $C = 0.0090$.		$a = 1.9713$; $\alpha = 0.8293$; $X^2_4 = 2.954$; $P = 0.5656$.	

Text 7: F. Weidenreich, Eine neu entdeckte Übergangsform zwischen dem Neandertaler und dem heutigen Menschen. *Natur und Museum* 62/1932: 384-389.
Text 8: F. Dewers, Die geologische Bedeutung der Pflanzenwurzeln. *Natur und Museum* 63/1933: 253-259.
Text 9: E. Henning, Wieder auf den Spuren der Schreckens-Echsen. *Natur und Volk* 64/1934: 322-324.

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	481	480.89	535	533.58	454	452.97
2	303	296.61	306	288.60	293	288.82
3	145	154.72	147	174.84	154	167.82
4	61	60.53	83	79.44	87	73.14
5	22	18.94	42	28.88	18	25.50
6	5	4.94	4	11.69	7	7.41
7	-	-	-	-	5	2.38
8	1	1.39	-	-	-	-
	$a = 1.5649$; $\alpha = 0.9000$; $X_4^2 = 1.343$; $P = 0.8540$;		$a = 1.8174$; $\alpha = 0.8066$; $X_2^2 = 6.381$; $P = 0.0411$; $C = 0.0057$.		$a = 1.7432$; $\alpha = 0.8806$; $X_2^2 = 4.610$; $P = 0.0997$.	

Text 10: P. Rietschel, Vom Honigtau. *Natur und Volk* 65/1935: 322-326.
Text 11: W.E. Ankel, Die Netz-Reusenschnecke, ein Aasfresser im Watt. *Natur und Volk* 66/1936: 341-345.
Text 12: P. Eipper, Menschen-Affen. *Natur und Volk* 68/1938: 310-315.

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	537	535.86	428	427.35	575	572.45
2	348	333.85	275	267.76	369	350.58
3	168	192.83	142	155.52	147	180.58
4	103	83.53	72	67.74	74	69.77
5	29	28.95	25	23.61	24	21.56
6	1	11.01	8	6.86	13	7.09
7	-	-	1	2.19	-	-
	$a = 1.7328$; $\alpha = 0.8731$; $X_1^2 = 4.536$; $P = 0.0332$; $C = 0.0038$		$a = 1.7424$; $\alpha = 0.8738$; $X_4^2 = 2.539$; $P = 0.6377$.		$a = 1.5453$; $\alpha = 0.9012$; $X_1^2 = 8.841$; $P = 0.0029$; $C = 0.0074$.	

Text 13: R. Bott, Die Lehmwespe *Odynerus parietum* und ihre eigenartige Vernichtung durch die Schmarotzerfliege *Meigenia floralis*. *Natur und Volk* 69/1939: 542-547.
Text 14: E. Franz, Die Gallen der Rosen-Gallwespe (*Rhodites rosae*). *Natur und Volk* 70/1940: 407-410.
Text 15: E. Nowack, Reste der ältesten Lebewelt des afrikanischen Festlandes gelangen in deutsche Museen. *Natur und Volk* 70/1940: 557-564.

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	438	436.30	351	349.94	500	498.85
2	259	247.99	230	219.78	331	325.80
3	107	129.63	142	152.62	175	191.54
4	60	50.82	79	79.48	98	84.45
5	18	15.94	33	33.12	31	29.79
6	4	5.34	8	11.50	6	8.76
7	-	-	8	4.60	0	2.21
8	-	-	-	-	1	0.63
	$a = 1.5682$; $\alpha = 0.8645$; $X_3^2 = 6.698$; $P = 0.0822$.		$a = 2.0832$; $\alpha = 0.8367$; $X_4^2 = 4.855$; $P = 0.3025$.		$a = 1.7637$; $\alpha = 0.8867$; $X_4^2 = 5.768$; $P = 0.2172$.	

Text 16: G. Eberle, Nacht-Pfauenauge und Birkenspinner, zwei Frühlingsboten. *Natur und Volk* 71/1941: 196-201.
Text 17: R. Duden, Der „Brotkrumeschwamm", ein Kieselschwamm der Nordsee. *Natur und Volk* 80/1950: 216-220.
Text 18: H. Schmidt, Ausgeprägte Form bei Tieren. *Natur und Volk* 81/1951: 84-89.

	Text 19	
x	n_x	NP_x
1	428	427.24
2	239	241.62
3	138	139.65
4	72	60.53
5	15	20.99
6	6	7.99
	$a = 1.7339$ $\alpha = 0.8346$; $X_3^2 = 4.418$; $P = 0.2197$.	

Text 19: R. Mertens, Der „Bärtige Krötenkopf" und seine Warnstellung. *Natur und Volk* 82/1952: 15-19.

6. In der folgenden Tabelle werden die relativen Häufigkeiten der Wortlänge n_x in % wiedergegeben.

x / Text	1	2	3	4	5	6	7	8	9	10	21
T 1	47,2	25,7	15,3	6,7	2,8	1,4	0,5	0,3	0,1		
T 2	45,9	24,8	15,5	9,0	4,1	0,6		0,1			
T 3	48,2	24,1	16,8	8,3	1,6	0,7	0,3				
T 4	45,9	23,2	18,6	7,2	3,3	1,2	0,4	0,2			
T 5	41,2	23,9	17,6	11,5	3,3	2,0	0,2	0,1	0,1		
T 6	45,5	26,4	14,9	9,8	1,7	0,9	0,5	0,2		0,1	
T 7	46,0	25,0	14,0	7,1	4,9	2,1	0,8	0,1			
T 8	42,5	26,9	15,8	10,3	3,6	0,7	0,1				
T 9	43,9	26,2	15,9	8,7	4,1	1,1	0,1	0,1			
T 10	47,2	29,8	14,2	6,0	2,2	0,5					0,1
T 11	47,9	27,4	13,2	7,4	3,8	0,4					
T 12	44,6	28,8	15,1	8,5	1,8	0,7	0,5				
T 13	45,3	29,3	14,2	8,7	2,4	0,1					
T 14	45,0	28,9	14,9	7,6	2,6	0,8	0,1				
T 15	47,8	30,7	12,2	6,2	2,0	1,1					
T 16	49,4	29,2	12,1	6,8	2,0	0,5					
T 17	41,2	27,0	16,7	9,3	3,9	0,9	0,9				
T 18	43,8	29,0	15,3	8,6	2,7	0,5		0,1			
T 19	47,7	26,6	15,4	8,0	1,7	0,7					

Alle Texte zeigen, daß einsilbige Wörter am häufigsten auftreten: die relativen Häufigkeiten nehmen mit steigender Silbenzahl ab. Bis auf eine Ausnahme (Extremfall mit 21 Silben: eine Ziffer in Text 10) treten größere Wortlängen (sieben und mehr Silben) gar nicht bzw. in sehr geringem Maße auf. Die mittleren Wortlängen (drei- und viersilbige Wörter) zeigen im Vergleich zu den bearbeiteten deutschsprachigen Kurzprosatexten (Best & Zhu, 1994: Tabelle Seite 27) größere Häufigkeiten; das ist wohl auf den fachwissenschaftlichen Funktionalstil zurückzuführen.

7. Die Untersuchung deutschsprachiger naturwissenschaftlicher Publikationen hat gezeigt, daß sich diese Texte mit der *positiven Singh-Poisson-Verteilung* modellieren lassen. Nur ein weiterer (W. Jacobs, Beobachtungen an der Heuschrecke *Calliptamus italicus*. *Natur und Volk* 79/949:89-92.) der insgesamt 20 untersuchten Texte konnte durch diese Verteilung nicht dargestellt werden. Für einen solchen Mißerfolg kommen verschiedenen Ursachen infrage, z.B. mehrfache Bearbeitung durch den Autor, Eingriffe der Redaktion, u.ä.

Die Hypothese einer gesetzmäßigen Verteilung von Wortlängen kann durch diese Untersuchung weiterhin bekräftigt werden. Da sich hier die positive negative Binomialverteilung nicht als geeignetes Modell bewährt hat, zeigt sich außerdem die Annahme bestätigt, daß unterschiedliche Textsorten und Funktionalstile derselben Sprache verschiedenen Modellen folgen können. Allerdings handelt es sich bei der *positiven negativen Binomialverteilung* und der *positiven Poisson-Verteilung* lediglich um Varianten eines gemeinsamen Grundmodells (vgl. Wimmer & Altmann, 1996).

Textkorpus

Natur und Museum. Bd. 59-63, 1929-1933. Selbstverlag der Senckenbergischen Naturforschenden Gesellschaft, Frankfurt/M.

Natur und Volk. Bd. 64-44, 1934-1936; 68-71, 1938-1941; 79-82, 1949-1952. . Selbstverlag der Senckenbergischen Naturforschenden Gesellschaft, Frankfurt/M.

Literatur

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In P. Schmidt (Hg.), *Glottometrika 15* (S. 147-157), Trier: WVT.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In diesem Band,

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Franz Steiner Verlag.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax*. Frankfurt/M.: Athenäum-Verlag.

Duden. Aussprachewörterbuch. Bd. 6 (1974). Mannheim-Wien-Zürich: Dudenverlag.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

Glottometrika 16, 1997, 88-97

Zur Wortlängenhäufigkeit im Schwedischen: Gunnar Ekelöfs Briefe

Stefan Ammermann, Malin Bengtson

0. Einleitung

Mit großem Erfolg[1] konnte die Wortlängenhäufigkeit in Texten unterschiedlichster Textsorten in verschiedensten Sprachen mit Wahrscheinlichkeitsverteilungen modelliert werden. Dies bestätigt die Annahme, daß sprachliche Phänomene bestimmten Gesetzmäßigkeiten unterliegen. Die vorliegende Arbeit stellt einen Beitrag zu dieser Thematik dar und untersucht Briefe im Schwedischen.

1. Zur Textauswahl

In dieser Arbeit werden Briefe des schwedischen Autors Gunnar Ekelöf (1907 - 1968) untersucht. Ekelöf, der durch häufige Reisen und mehrmalige Wohnortwechsel auf die Briefkorrespondenz angewiesen war, stellt sich dem Leser in der ersten repräsentativen Ausgabe seiner Briefe aus dreierlei Sicht dar:

Die ersten Briefe - ab 1916 - schreibt der junge Ekelöf, der „rührende Briefe aus dem Kinderpensionat an die Mutter schreibt“[2]; von seinen Bildungsreisen aus England und Frankreich berichtet er als Heranwachsender seinen Freunden und Verwandten; in den späten Briefen schreibt schließlich der Poet und das Akademiemitglied Ekelöf. Für diese Untersuchung wurden Briefe aus der Zeit zwischen 1925 und 1935 ausgewählt. Es finden sich damit vorwiegend Briefe von seinen Bildungsreisen. So ist es auch nicht verwunderlich, daß hin und wieder ein englischer Name oder ein französisches Wort auftaucht.

2. Zur Textbearbeitung

Brieftexte sind für die Modellierung von Wortlängen eine besonders geeignete Textsorte, da sie meist spontan und ohne größere zeitliche Unterbrechung geschrieben werden und offenbar deshalb homogener wirken als andere Textsorten. Außerdem findet in der Regel keine Nachbearbeitung statt, wie dies bei anderen Textsorten üblich ist. Dadurch lassen sich eindeutigere Aussagen über den natürlichen Wortlängenrhythmus eines Autors treffen.

[1] vgl. dazu u.a.: Glottometrika 15, 16.

[2] Einleitung zum Briefband im Einband (Übers.: M. Bengtson).

Zur Untersuchung der Wortlängenhäufigkeit bieten sich verschiedene Möglichkeiten an. Dem ‚sprachlichen (linguistischen) Konstrukt‘ „Wort“ könnten demnach die ‚Konstituenten‘ „Morphem“, „Semantem“ oder „Silbe“ zugeordnet werden (Altmann & Schwibbe, 1989:8). Die Anzahl der jeweiligen ‚Konstituenten‘ im Wort bestimmt dessen Länge. Es hat sich in der Praxis der Wortlängenanalyse herausgestellt, daß die Silbe als leicht definierbare und somit „problemlose“ ‚Konstituente‘ den anderen vorzuziehen ist.

Nachfolgend sollen die Einheiten „Wort“ und „Silbe“ für diese Untersuchung näher bestimmt werden. Für „Wort“ wurde die Definition des „orthographischen Wortes“ zugrunde gelegt (Lühr, 1993:131f). Der Bindestrich betont die Einheit des Wortes, das Apostroph weist auf die Worttrennung hin. Letzteres kam in der vorliegenden Untersuchung aber nicht vor und sei deshalb hier nur der Vollständigkeit halber erwähnt. Die Anzahl der Silben richtet sich nach der Anzahl der Vokale und Diphthonge im Wort.

Nachfolgend sei noch auf einige Entscheidungen bei der Datenerhebung hingewiesen:

Gewertet wurde jeweils nur der laufende Text, d.h. Anrede, Ortsangabe und Datum sowie die Unterschrift sind nicht als dem Text zugehörig betrachtet worden. Abkürzungen wurden in ihrer ausformulierten Form gezählt. Zahlwörter wurden in ihrer mündlich realisierten Form gewertet und nach Tausendern, Hundertern und Zehnern in Einzelwörter zerlegt. Demnach ist „32“ als ein Wort mit vier Silben, „332“ aber als zwei Wörter mit drei und vier Silben zu zählen. Auch Jahreszahlen würden auf diese Weise gewertet, kamen aber in den untersuchten Briefen nicht vor. Das Wortbildungsmorphem „-ion“ z.B. in „Religion“ (Text 1) wird im Schwedischen phonologisch als / u: n / realisiert und ist somit einsilbig. Das Wort Religion ([reli(j)u:n]) ist aus diesem Grund als dreisilbiges Wort gewertet.

Die französischen Wörter, die vereinzelt in den Briefen vorkommen, sind gemäß ihrer Aussprache gezählt worden.

3. Zur Problematik der Modellierung

Die bisherigen Untersuchungen von Brieftexten unterschiedlicher Autoren und aus verschiedenen Jahrhunderten haben keine einheitliche Festlegung auf ein Verteilungsmodell erbracht. Zwar lassen sich deutsche Brieftexte des 20. Jahrhunderts in der Regel mit der positiven negativen Binomialverteilung gut modellieren, aber schon Briefe des 19. Jahrhunderts zeigen diese Tendenz nicht mehr.

Für schwedische Briefe liegen bisher keine Untersuchungen vor. Es ist aber damit zu rechnen, daß in verschiedenen Sprachen, aber auch innerhalb einer Sprache, in verschiedenen Textsorten und Zeiten unterschiedliche Verteilungsmodelle verwendet werden müssen.

Wir gehen hier zunächst von Bests Untersuchung schwedischer Zeitungstexte (Best, 1996) aus, in der festgestellt wurde, daß die positive Singh-Poisson-Verteilung ein gutes Modell zumindest für diese Textsorte darstellt. Allerdings muß erst noch geprüft werden, ob sich dieses Modell für das Schwedische der Gegenwart insgesamt bewährt oder ob hier mit unterschiedlichen Modellen zu rechnen ist. Den theoretischen Rahmen dafür haben Wimmer u.a. (1994) sowie Wimmer & Altmann (1996) vorgegeben; in diesen Arbeiten wird das Spektrum theoretisch begründbarer Wortlängenverteilungsmodelle entwickelt; auch wenn es sich dabei nur um einen Ausschnitt der möglichen Modelle handelt, wollen wir unsere Versuche, Verteilungen an die gefundenen Daten anzupassen, auf die dort genannten beschränken.

Bei der Prüfung verschiedener Modelle hat sich nun gezeigt, daß die positive Singh-Poisson-Verteilung auch an die Briefe Ekelöfs angepaßt werden konnte; diese Verteilung kann damit weiterhin als ein gutes Modell für Texte des zeitgenössischen Schwedisch angesehen werden. In den Tabellen zu den einzelnen Briefen sind deshalb unter NP_x die berechneten Werte der positiven Singh-Poisson-Verteilung angegeben, deren Formel wie folgt lautet:

$$P_x = \begin{cases} 1-\alpha+\dfrac{\alpha\, a\, e^{-a}}{1-e^{-a}}, & x = 1 \\[2ex] \dfrac{\alpha\, a^x e^{-a}}{x!\,(1-e^{-a})}, & x = 2,3,4,\ldots \end{cases}$$

Unsere Berechnungen haben darüber hinaus aber gezeigt, daß auch weitere von den in Frage kommenden Verteilungen mit z.T. größerem Erfolg an die Brieftexte angepaßt werden konnten. Besonders gute Ergebnisse erbrachte die Anwendung der 1-verschobenen Hyperpoisson-Verteilung, die wir deshalb unter NP_{x1} in den Tabellen ebenfalls angeben, deren Formel wie folgt lautet:

$$P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1\,(1;b;a)}, \quad x = 1,2,3,\ldots$$

wo ${}_1F_1$ (.) die hypergeometrische Funktion darstellt.

Nur der Vollständigkeit halber sei erwähnt, daß bei weiteren Untersuchungen zu schwedischen Texten der Gegenwart auch die Consul-Jain-Poisson-Verteilung und die positive negative Binomialverteilung in Betracht gezogen werden sollten, die sich an die Brieftexte Ekelöfs ebenfalls mit Erfolg anpassen ließen, wenn auch nicht ganz so gut wie die Hyperpoisson-Verteilung. Es wird deshalb hier darauf verzichtet, auch deren Anpassungen im Einzelnen mitzuteilen.

4. Untersuchungsergebnisse

Als bestes Modell hat sich die Hyperpoisson-Verteilung herausgestellt. Die folgenden Tabellen zeigen die Ergebnisse der Anpassung der positiven Singh-Poisson-Verteilung (NP_x) und der Hyperpoisson-Verteilung (NP_{x1}). Dabei werden die jeweiligen Werte für die Parameter a, b und α, das Chiquadrat (χ^2), die Anzahl der Freiheitsgrade (als Index im Chiquadrat), die Überschreitungswahrscheinlichkeit des Chiquadrats (P) und die Kontingenz (C), falls notwendig, angegeben.

	Text 1			Text 2		
x	n_x	NP_x	NP_{x1}	n_x	NP_x	NP_{x1}
1	121	120,88	120,82	163	162,47	164,72
2	64	62,68	64,16	79	76,89	76,42
3	27	29,37	27,67	28	32,78	30,18
4	11	10,32	10,05	13	10,48	10,38
5	3	2,90	3,15	2	2,68	3,16
6	1	0,84	1,14	0	0,57	0,86
7	-	-	-	1	0,12	0,27
$\sum$	227			286		
		$a = 1{,}41$;	$a = 2{,}30$;		$a = 1{,}28$;	$a = 2{,}66$;
		$\alpha = 0{,}86$;	$b = 4{,}33$;		$\alpha = 0{,}85$;	$b = 5{,}73$;
		$X_2^2 = 0{,}28$;	$X_3^2 = 0{,}13$;		$X_2^2 = 1{,}40$;	$X_3^2 = 1{,}37$;
		$P = 0{,}87$.	$P = 0{,}99$.		$P = 0{,}50$.	$P = 0{,}71$.

Text 1 „Till Hanna von Hedenberg“ vom 24.6.1925
Text 2 „Till Valborg Hahr“ vom 28.5.1926

	Text 3			Text 4		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	190	189,96	189,72	193	201,34	192,7
2	107	106,15	106,84	110	94,72	108,24
3	42	43,98	43,26	26	30,73	29,14
4	15	13,67	13,67	6	7,48	5,16
5	2	3,40	3,54	1	1,73	0,76
6	1	0,70	0,78	-	-	-
7	1	0,15	0,18	-	-	-
Σ	358			336		
		$a = 1,24$;	$a = 1,44$;		$a = 0,97$;	$a = 0,52$;
		$\alpha = 0,95$;	$b = 2,56$;		$\alpha = 0,98$;	$b = 0,92$;
		$X_2^2 = 0,24$;	$X_2^2 = 0,22$;		$X_2^2 = 4,14$;	$X_1^2 = 0,57$;
		$P = 0,89$.	$P = 0,90$.		$P = 0,13$.	$P = 0,45$.

Text 3 „Till Sara Hahr“ vom 2.7.1926
Text 4 „Till Hanna von Hedenberg“ vom 6.9.1926

	Text 5			Text 6		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	198	197,06	199,16	103	102,02	104,63
2	115	108,62	106,57	73	66,43	61,37
3	34	43,43	42,37	23	31,84	30,2
4	13	13,02	13,4	10	11,44	12,8
5	4	3,12	3,52	3	3,29	4,77
6	1	0,62	0,79	1	0,79	1,58
7	0	0,11	0,15	2	0,16	0,47
8	1	0,02	0,03	1	0,03	0,17
Σ	366			216		
		$a = 1,20$;	$a = 1,55$;		$a = 1,44$;	$a = 3,06$;
		$\alpha = 0,96$;	$b = 2,90$;		$\alpha = 0,96$;	$b = 5,22$;
		$X_2^2 = 3,59$;	$X_2^2 = 2,84$;		$X_2^2 = 5,03$;	$X_3^2 = 6,63$;
		$P = 0,17$.	$P = 0,24$.		$P = 0,08$.	$P = 0,09$.

Text 5 „Till Carl Nordenfalk“ vom 31.5.1928
Text 6 „Till Carl Nordenfalk“ vom 12.10.1928

	Text 7			Text 8		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	146	145,68	146,45	419	415,01	423,87
2	75	73,59	74,54	214	206,86	197,49
3	32	32,91	30,87	54	71,20	69,36
4	9	11,04	10,77	22	18,38	19,55
5	4	2,96	3,25	7	4,56	5,73
6	0	0,66	0,86	-	-	-
7	1	0,15	0,26	-	-	-
Σ	267			716		
		$a = 1,34$;	$a = 2,22$;		$a = 1,03$;	$a = 1,42$;
		$\alpha = 0,87$;	$b = 4,37$;		$\alpha = 0,98$;	$b = 3,06$;
		$X_2^2 = 0,83$;	$X_3^2 = 0,52$;		$X_2^2 = 6,46$;	$X_2^2 = 5,43$;
		$P = 0,66$.	$P = 0,91$.		$P = 0,04$.	$P = 0,07$.

Text 7 „Till Gunnar Hahr“ vom 24.12.1928
Text 8 „Till Valborg Hahr“ vom 21.1.1929

	Text 9			Text 10		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	189	188,32	188,44	175	176,00	175,74
2	87	82,80	85,31	94	84,95	85,28
3	31	37,12	33,9	19	28,30	28,20
4	12	12,48	12,01	7	7,07	7,07
5	5	3,36	3,83	3	1,69	1,71
6	1	0,93	1,51	-	-	-
Σ	325			298		
		$a = 1,35$;	$a = 3,25$;		$a = 0,99$;	$a = 1,04$;
		$\alpha = 0,80$;	$b = 7,18$;		$\alpha = 0,98$;	$b = 2,14$;
		$X_2^2 = 1,93$;	$X_3^2 = 0,81$;		$X_2^2 = 5,05$;	$X_2^2 = 4,87$;
		$P = 0,38$.	$P = 0,85$.		$P = 0,08$.	$P = 0,09$.

Text 9 „Till Sara och Alfred Hahr“ vom 7.11.1929
Text 10 „Till Valborg Hahr“ vom 6.2.1930

	Text 11			Text 12		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	214	219,52	214,17	177	185,82	177,69
2	111	102,59	112,26	118	103,47	115,67
3	34	33,06	31,53	37	39,89	39,58
4	5	7,99	6,05	10	11,53	9,19
5	1	1,83	1,00	1	2,67	1,61
6	-	-	-	1	0,61	0,26
Σ	365			344		
		$a = 0,97$;	$a = 0,6$;		$a = 1,1$;	$a = 0,7$;
		$\alpha = 0,9$;	$b = 1,1$;		$\alpha = 0,9$;	$b = 1,1$;
		$X_2^2 = 2,3$;	$X_1^2 = 0,3$;		$X_2^2 = 3,3$;	$X_2^2 = 0,3$;
		$P = 0,31$.	$P = 0,55$.		$P = 0,19$.	$P = 0,86$.

Text 11 „Till Gunnel Bergström“ vom 15.8.1931
Text 12 „Till Gunnel Bergström“ vom 23.8.1931

	Text 13			Text 14		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	62	60,86	60,54	324	325,98	326,3
2	31	28,36	32,55	157	159,24	149,53
3	13	19,85	16,89	59	62,24	61,33
4	15	10,41	8,46	23	18,25	22,77
5	3	4,37	4,1	3	4,28	7,72
6	1	1,53	1,92	4	0,84	2,41
7	0	0,46	0,87	1	0,16	0,94
8	0	0,12	0,39	-	-	-
9	1	0,04	0,28	-	-	-
Σ	126			571		
		$a = 2,10$;	$a = 14,71$;		$a = 1,17$;	$a = 3,90$;
		$\alpha = 0,73$;	$b = 27,35$;		$\alpha = 0,91$;	$b = 8,53$;
		$X_3^2 = 5,09$;	$X_4^2 = 6,98$;		$X_2^2 = 2,85$;	$X_3^2 = 4,18$;
		$P = 0,17$.	$P = 0,14$.		$P = 0,24$.	$P = 0,24$.

Text 13.„Till Verner von Heidenstam“ vom 29.9.1932
Text 14 „Till Elmer Diktonius“ vom 23.12.1932

	Text 15			Text 16		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	149	146,16	146,89	254	251,37	252,53
2	85	74,92	82,40	172	159,95	160,59
3	33	52,35	44,35	58	82,13	78,77
4	31	27,44	22,94	40	31,63	31,44
5	17	11,50	11,42	12	9,74	10,58
6	3	5,63	10	2	3,18	4,09
Σ	318			538		
		$a = 2,10$;	$a = 13,30$;		a = 1,54;	a = 2,15;
		$\alpha = 0,77$;	$b = 23,71$;		$\alpha = 0,92$;	b = 3,37;
		$X_3^2 = 12,88$;	$X_3^2 = 5,94$;		$X_3^2 = 11,20$;	$X_3^2 = 9,88$;
		$P = 0,005$;	$P = 0,05$.		$P = 0,01$;	$P = 0,02$;
		$C = 1,6^{-5}$			$C = 2^{-5}$.	$C = 4^{-5}$.

Text 15 „Till Kaj Bonnier“ vom 26.12.1932
Text 16 „Till Rabbe Enckell“ vom 20.1.1933

	Text 17			Text 18		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	130	125,86	128,61	185	184,90	185,10
2	69	67,32	66,72	96	95,29	95,65
3	25	35,64	31,50	35	36,42	35,70
4	19	14,15	13,65	11	10,44	10,43
5	3	4,50	5,46	2	2,39	2,50
6	3	1,53	3,07	1	0,54	0,62
Σ	249			330		
		$a = 1,59$;	$a = 5,26$;		a = 1,15;	a = 1,35;
		$\alpha = 0,84$;	$b = 10,13$;		$\alpha = 0,94$;	b = 2,60;
		$X_3^2 = 6,94$;	$X_3^2 = 4,65$;		$X_2^2 = 0,09$;	$X_2^2 = 0,05$;
		$P = 0,07$.	$P = 0,20$.		$P = 0,96$.	$P = 0,98$.

Text 17 „Till Johannes Edfelt“ vom 30.7.1935
Text 18 „Till Knut Jaensson“ vom 13.8.1935

	Text 19			Text 20		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
1	79	77,61	79,09	317	310,90	320,77
2	33	30,61	27,11	142	136,95	128,18
3	5	10,45	9,71	33	51,56	46,69
4	3	2,68	2,69	19	14,56	15,63
5	1	0,55	0,83	5	3,29	4,84
6	1	0,11	0,30	2	0,74	1,89
$\sum$	123			518		
	$a = 1{,}20$; $\alpha = 0{,}7$; $X_1^2 = 4{,}8$; $P = 0{,}0$; $C = 0{,}0404$.		$a = 20657{,}04$ $b = 57214{,}63$ $X_1^2 = 3{,}97$; $P = 0{,}14$.		$a = 1{,}13$; $\alpha = 0{,}87$; $X_2^2 = 10{,}54$; $P = 0{,}005$; $C = 9{,}7^{-6}$.	$a = 4{,}12$; $b = 10{,}31$; $X_3^2 = 6{,}29$; $P = 0{,}10$.

Text 19 „Till Maj Strindberg" vom 17.8.1935
Text 20 „Till Hager Olsson" vom 26.12.1935

5. Bewertung der Ergebnisse

Es kann festgestellt werden, daß sich alle 20 Briefe Gunnar Ekelöfs mit der Hyperpoisson-Verteilung modellieren lassen. Die berechneten P-Werte entsprechen bei 19 Briefen dem Kriterium von $P \geq 0{,}05$. Lediglich bei Text 16 mußte zusätzlich der C-Wert betrachtet werden, um die Anpassung zu bestätigen. Auffällig hierbei war, daß dieser Brief bei allen 18 geprüften Verteilungsmodellen schlechte P-Werte lieferte. Die Gründe hierfür können in stilistischen Besonderheiten des Textes liegen, die zu dieser auffälligen Abweichung führen. Text 19 weist sehr große Parameter a und b der Hyperpoisson-Verteilung auf. Offensichtlich vollzieht sich hier die Konvergenz zu der geometrischen Verteilung ($a \to \infty$, $b \to \infty$, $a/b \to q$). In der Tat kann man zeigen, daß die 1-verschobene geometrische Verteilung eine sehr gute Anpassung liefert ($X_3^2 = 3.96$, $P = 0.27$).

Best (1996) konnte zeigen, daß schwedische Pressetexte in der Regel der positiven Singh-Poisson-Verteilung folgen. Mit Hinweis auf eine sich möglicherweise verändernde Strukturierung dieser Textsorte im Schwedischen nennt er außerdem die Consul-Jain-Poisson-Verteilung als besseres Modell für solche Texte, die von dieser Regel abweichen.

Auch die Briefe Ekelöfs lassen sich mit der positiven Singh-Poisson-Verteilung modellieren. Die Anpassungen sind aber nicht ganz so gut wie die durch die Hyperpoisson-Verteilung. Eine auffällige Parallele bieten Beobachtungen an deutschen Brieftexten des 16. bis 19. Jhds.: Die Untersuchung von je 20 Briefen aus jedem der Jahrhunderte hat gezeigt, daß von den insgesamt 80 Texten 79 der Hyperpoisson-Verteilung folgen (Ammermann, 1996). Es könnte sein, daß die Hyperpoisson-Verteilung sich zur Modellierung dieser Textsorte als besonders geeignet erweist.

Die Beobachtung, daß verschiedene Modelle zufriedenstellende Anpassungsergebnisse liefern können, konnte in dieser Arbeit ebenfalls bestätigt werden. Es müssen nun noch die Gründe für diese Mehrfachmodellierung aufgedeckt werden, indem die Ergebnisse textnah bewertet werden und so unter ihren Randbedingungen erscheinen. Es liegt dabei nahe, die Unterschiede bei der Modellierung der Briefe und der Pressetexte dem Einfluß der Textsorten oder der Autoren zuzuschreiben.

Literatur

Altmann, G., & Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Altmann, G. (1994). *Altmann-Fitter: Iterative Anpassung diskreter Wahrscheinlichkeitsverteilungen*. 1. Aufl., Lüdenscheid: RAM-Verlag.

Ammermann, S. (1996). Wortlängen in deutschen Briefen seit fnhd. Zeit. In Arbeit.

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In P. Schmidt (Hg.), *Glottometrika 15* (S. 147-157), Trier: WVT.

Ekelöf, G. (1989) Brev 1916 - 1968. Urval och kommentarer Carl Olov Sommar, Stockholm: Bonnier.

Lühr, R. ([4]1993). *Neuhochdeutsch*. München: Fink.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1, 98 - 106.

Wimmer, G., & Altmann, G. (1996). The theory of word length distribution: Some results and generalizations. In P. Schmidt (Hg.), *Glottometrika* 15 (S. 112-133), Trier: WVT.

Glottometrika 16, 1997, 98-107

Zur Häufigkeit von Wortlängen in englischen Texten

Alice Hasse, Michaela Weinbrenner

0. Allgemeines

Im Rahmen eines Projekts an der Universität Göttingen 1993/94 zur Untersuchung der Häufigkeit von Wortlängen verschiedener Sprachen wurden von uns englische Texte auf ihre Wortlängenhäufigkeit hin untersucht. Dabei handelte es sich zum einen um Briefe von Schriftstellern aus der ersten Hälfte des 20. Jahrhunderts und zum anderen um Artikel aus aktuellen biologischen Fachzeitschriften.

Die Wortlängen in den einzelnen Texten wurden in Silben gemessen. wobei folgende operationale Definitionen für *Wort* und *Silbe* zugrundegelegt wurden: Als *Wort* wurde die „orthographische Einheit" angesehen. Die *Silbenzahl eines Wortes* wurde aus der Anzahl der im Wort vorkommenden Vokale bzw. Diphthonge bestimmt. Abkürzungen und Symbole wurden entsprechend ihrer mündlich realisierten Form ausgewertet.[1]

Die Markierung des Genitivs im Englischen durch *'s* wurde in allen Texten als zum Wort gehörig behandelt. Bei den Briefen wurden Triphthonge als zwei Silben gewertet; bei den biologischen Fachtexten wurden sie hingegen als ein Laut (eine Silbe) angesehen. Diese Unterschiede bei der Handhabung der Daten wurden im nachhinein nicht verändert. da sie sich bei den Rechnungen nicht störend bemerkbar gemacht haben. Des weiteren liegt den biologischen Artikeln die *received pronunciation* des britischen Englisch zugrunde. während die Briefe der amerikanischen Aussprache folgend bearbeitet wurden.[2]

Das auf diese Weise gewonnene Datenmaterial wurde daraufhin analysiert, ob es der gemischten Poisson-Verteilung in 1-verschobener Form mit folgender Formel entspricht:

[1] Nähere Ausführungen zu den Analyseprinzipien finden sich bei Best & Zhu (1994:20).

[2] Daraus resultiert teilweise ein Unterschied in der Silbenzahl: z.B. wird das Wort *Caribbean* im britischen Englisch viersilbig gesprochen. im amerikanischen Englisch dagegen dreisilbig (s. Text 17).

$$P_x = \frac{\alpha e^{-a} a^{x-1}}{(x-1)!} + \frac{(1-\alpha) e^{-b} b^{x-1}}{(x-1)!}, \qquad x = 1, 2, 3, \ldots$$

Die Güte der Anpassung wurde mit dem Chiquadrat-Test überprüft: Sie wurde als zufriedenstellend betrachtet, wenn $P(X^2) \geq 0.05$ bzw. $C \leq 0.02$ ist. Den Koeffizienten C benutzt man üblicherweise bei großem Stichprobenumfang oder - notgedrungen - in solchen Fällen, in denen beim Testen keine Freiheitsgrade übrigbleiben.

1. Die Texte

A. Briefe:

Briefe als Textsorte zeichnen sich durch einen hohen Grad an thematischer Geschlossenheit und Spontaneität aus; sie haben meist nur einen Autor, der den Text ohne Unterbrechung verfaßt hat.

Ausgewählt wurden 16 Privatbriefe von drei amerikanischen und einem englischen Schriftsteller, deren Umfang zwischen 86 und 878 Wörtern liegt. In diesen Texten wird fast ausnahmslos der Funktionalstil der Alltagssprache verwendet, selbst wenn es um Themen wie Literatur oder Politik geht.

Ein typisches Merkmal des alltagssprachlichen Funktionalstils bildet sowohl das Auftreten von nullsilbigen Wörtern, die durch Enklise entstanden sind (z.B. *I've* aus *I have*, *I'm* aus *I am* etc.)[3], als auch von kontrahierten Verneinungen (z.B. *can't* aus *cannot*, *don't* aus *do not* etc.), die in den wissenschaftlichen Texten nicht auftreten (vgl. B.). Die durch Enklise entstandenen Formen wurden ebenso wie Verneinungen mit Kontraktion als **ein** Wort gewertet; im letzteren Fall beschreibt die Zusammenziehung nicht Morphemgrenzen, sondern wird innerhalb des Morphems *not* vorgenommen.

Aufgrund mangelnder Einheitlichkeit wurden die Briefköpfe nicht berücksichtigt; es wurde also immer nur der laufende Text ausgezählt.

Die Ergebnisse:

In den Tabellen bedeuten: x Wortlänge in Silben, n_x die beobachtete Häufigkeit, NP_x die nach der gemischten Poisson-Verteilung berechnete Häufigkeit, a, b und α sind Parameter, X^2 ist das Chiquadrat, P die Überschreitungswahrscheinlichkeit des Chiquadrats, $C = X^2/N$ der Diskrepanzkoeffizient.

[3] Das Auftreten von nullsilbigen Enklitika in englischen Texten hängt allein von pragmatischen (also außersprachlichen) Faktoren ab, z.B. dem Formalitäts- und Spontaneitätsgrad eines Textes; d.h. ihr Vorkommen ist nicht von quantitativen Relationen bestimmt.

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	249	248.92	273	272.73	59	58.79
2	58	57.97	51	50.62	17	17.53
3	15	15.36	9	10.23	7	6.62
4	6	4.95	4	3.17	2	2.06
5	0	1.39	1	1.25	-	-
6	1	0.41	-	-	-	-
	$a = 1.1532$; $b = 0.1577$; $\alpha = 0.1810$; $X_1^2 = 0.57$; $P = 0.45$.		$a = 1.2299$; $b = 0.1468$; $\alpha = 0.0991$; $X_1^2 = 0.41$; $P = 0.52$.		$a = 0.7597$; $b = 0.0027$; $\alpha = 0.5771$; $X_0^2 = 0.04$; $C = 0.0005$.	

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	346	344.79	291	291.01	134	133.70
2	64	64.24	72	72.14	45	45.92
3	28	29.12	19	18.70	11	9.76
4	12	12.90	4	4.25	1	2.03
5	7	5.95	1	0.90	1	0.59
	$a = 1.3660$; $b = 0.0726$; $\alpha = 0.2601$; $X_1^2 = 0.30$; $P = 0.58$.		$a = 0.7239$; $b = 0.1069$; $\alpha = 0.3544$; $X_0^2 = 0.01$; $C = 0.00003$.		$a = 1.1553$; $b = 0.3090$; $\alpha = 0.0901$; $X_0^2 = 0.32$; $C = 0.002$.	

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	317	317.30	252	251.96	293	293.03
2	82	82.16	64	63.92	67	66.92
3	27	26.34	17	17.46	20	20.25
4	10	9.48	5	4.52	8	7.59
5	2	2.84	1	1.14	2	2.41
6	0	0.69	-	-	1	0.80
7	1	0.19	-	-	-	-
	$a = 1.2212$; $b = 0.1561$; $\alpha = 0.2367$; $X_1^2 = 0.18$; $P = 0.67$.		$a = 0.8479$; $b = 0.1288$; $\alpha = 0.3015$; $X_1^2 = 0.08$; $P = 0.78$.		$a = 1.2929$; $b = 0.1475$; $\alpha = 0.1927$; $X_1^2 = 0.04$; $P = 0.85$.	

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	646	645.42	286	283.63	620	619.54
2	151	151.41	111	110.64	160	159.83
3	58	58.01	32	37.22	46	47.56
4	17	17.97	15	10.58	15	13.97
5	6	5.19	1	2.93	4	4.10
	$a = 0.9429$; $b = 0.0585$; $\alpha = 0.3758$; $X_1^2 = 0.19$; $P = 0.67$.		$a = 0.9198$; $b = 0.1818$; $\alpha = 0.4512$; $X_1^2 = 3.85$; $P = 0.05$.		$a = 0.9648$; $b = 0.1349$; $\alpha = 0.2854$; $X_1^2 = 0.13$; $P = 0.72$.	

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	173	172.59	377	376.63	166	164.92
2	33	33.53	78	77.82	34	34.22
3	12	12.19	22	23.56	10	11.99
4	4	3.69	9	7.55	6	4.87
5	-	-	2	2.44	-	-
	$a = 0.7429$; $b = 0.0054$; $\alpha = 0.4185$; $X_0^2 = 0.04$; $C = 0.0002$.		$a = 1.0382$; $b = 0.1075$; $\alpha = 0.2321$; $X_1^2 = 0.46$; $P = 0.50$.		$a = 0.9331$; $b = 0.0720$; $\alpha = 0.3043$; $X_0^2 = 0.61$; $C = 0.003$.	

Text 16:

x	n_x	NP_x
1	327	326.54
2	66	66.21
3	25	25.26
4	8	8.35
5	3	2.64
	$a = 1.0132$; $b = 0.0626$; $\alpha = 0.3092$; $X_1^2 = 0.07$; $P = 0.79$.	

Text 1: Robert Frost an George F. Whicher. 21 June 1918, Franconia[4].
Anmerkung: „agoraphobia" wurde als sechssilbiges Wort gewertet.
Text 2: Robert Frost an George Elliott. 21 January, 1922. Ann Arbor.
Anmerkung: „sociability" wurde als fünfsilbiges Wort gewertet.
Text 3: Robert Frost an Louis Untermeyer. December 23, 1922. South Shaftsbury[5].
Anmerkungen: Bei „adamant" handelt es sich um ein dreisilbiges französisches Wort. „R.F." wurden entsprechend ihrer phonologischen Realisierung als Buchstaben gewertet.

Text 4: Robert Frost an Louis Untermeyer. November 23, 1921. Ann Arbor.
Anmerkungen: „weren't" wurde als zweisilbiges, „temperature" als dreisilbiges Wort gewertet. Frost schreibt hier einen leicht ironischen Brief, den er mit Wortspielen anfüllt; das führt zum vermehrten Gebrauch vielsilbiger, sonst seltener Wörter (z.B. „anabolical", „diabolical", „katabolical", etc.).
Text 5: F. Scott Fitzgerald an Zelda Fitzgerald, May 11, 1940, Encino.
Anmerkungen: „Russian" und „really" wurden als zweisilbige Wörter angesehen. Die Abkürzung „etc." ist aus zwei Wörtern entstanden und dementsprechend als *et* (einsilbig) und *cetera* (dreisilbig) gezählt worden; „1920" wird folgendermaßen ausgeschrieben: *nineteen hundred twenty* und wurde entsprechend gewertet.
Text 6: F. Scott Fitzgerald an Ernest Hemingway, April 18, 1927, Ellerslie.
Anmerkungen: „$ 200.00" wurde als *two hundred dollars* gelesen und „1st" als *first* und entsprechend dieser Auflösung gewertet.

Text 7: F. Scott Fitzgerald an Ernest Hemingway, November 1927, Ellerslie.
Anmerkungen: „obviously", „undistinguished" und „enthusiasm" wurden als viersilbige Wörter gewertet. „7500" wurde als *seven thousand five hundred* und „$ 3500" als *three thousand five hundred dollars* gelesen. Siebensilber: „individuality".
Text 8: F. Scott Fitzgerald an Zelda Fitzgerald, October 11, 1940, Hollywood.
Anmerkung: „Riviera" wurde als dreisilbiges Wort behandelt.
Text 9: George Orwell an Victor Gollacz. 9 May 1937, Barcelona.
Anmerkungen: „P.O.U.M." ist die Abkürzung für *Partido Obrero de Unificación Marxista* und wurde entsprechend der Aussprache der Buchstaben gewertet, bei „A." (für *Arthur*) und „B.C." (für *Book club*) wurde ebenso verfahren.

[4] Text 1 und 2 wurden aus Robert Frost (1964) entnommen.

[5] Text 3 und 4 wurden aus Robert Frost (1963) entnommen.

Text 10: George Orwell an R. Heppenstall, 31 July 1937, Wallington.
An-merkungen: „Margaret" wurde als zweisilbiges Wort, „interesting", „physically" und „practically" als dreisilbige Wörter gewertet. actually" dagegen als viersilbiges. „Xmas" bildet eine unkonventionelle Abkürzung von *Christmas*, einem zweisilbigen Wort. „Au revoir" wurde entsprechend seiner französischen Aussprache behandelt.
Text 11: George Orwell an Alec H. Joyce, 12 February 1938, Wallington.
Anmerkungen: „Indian" wurde als zweisilbiges, „Imperial" als viersilbiges Wort behandelt. „P.S." wurde als zwei einsilbige Wörter gewertet. Dieser Brief enthält einen kurzen Lebenslauf Orwells, aus diesem Grund finden sich hier sehr viele Jahreszahlen, was sich v.a. bei den ein-, zwei- und dreisilbigen Wörtern bemerkbar macht: „1903" gelesen als *nineteen hundred and three*, „1917-1921" als *from nineteen hundred seventeen to nineteen hundred twenty-one*, „1928-9" gelesen als *from nineteen hundred twenty-eight to nine*; weitere Jahreszahlen: „1902", „1933", „1936", „1937" und"1922-1927".
Text 12: George Orwell an John Sceats, 26 October 1938, Gueliz.
Anmerkungen: „isn't" wurde als zweisilbiges Wort gewertet, „L.P." als zwei einsilbige Wörter. „£ 5" wird als *five pounds* gelesen und „50,000" als *fifty thousand*. „Czechoslovakia" bildet ein fünfsilbiges Wort.

Text 13: William Carlos Williams an Harriet Monroe.
Anmerkungen: Der Buchstabe „W." wird als dreisilbiges Wort ausgesprochen, der Buchstabe „C." hingegen als Einsilber.
Text 14: William Carlos Williams an Alva N. Turner, June 25, 1919.
Anmerkungen: „genius" wurde als zweisilbiges Wort gewertet, „1" übernimmt die Funktion von *first*, einem einsilbigen Wort.
Text 15: William Carlos Williams an Marianne Moore, September 26, 1931.

Text 16: William Carlos Williams an Ezra Pound, March 15, 1933.
Anmerkungen: „mechanism" wurde als dreisilbiges, „immediately" als viersilbiges Wort gewertet. „shd." erscheint einmal als Abkürzung von *should*, „&" als Abkürzung von *and* und „Yrs." als Abkürzung von *Yours*, diese wurden entsprechend ihrer phonologischen Realisierung gewertet.

B. Artikel aus biologischen Fachzeitschriften:
Im Unterschied zu Briefen werden wissenschaftliche Publikationen häufig von Autorenkollektiven verfaßt. Daraus und aus den konventionellen und inhaltlichen Vorgaben (z.B. vorgegebene Gliederungspunkte), die dieser Textsorte eigen sind, ergibt sich, daß sie selten spontan und ohne Unterbrechung verfaßt sind. Zudem sind die Autoren zum Teil Nichtmuttersprachler, die in internationalen Zeitschriften auf Englisch veröffentlichen. Da dies sicherlich in einigen Fällen durch Übersetzung eines in ihrer jeweiligen Muttersprache abgefaßten Textes geschieht, könnten sich Abweichungen von anderen englischen Texten ergeben.

Es wurden fünf Artikel aus aktuellen biologischen Fachzeitschriften ausgewählt, die einen Umfang von 604 bis 1.738 Wörtern aufweisen. Sie sind im Funktionalstil der Wissenschaft verfaßt, so daß sie in der Regel keinerlei Enklitika enthalten. Lediglich der populärwissenschaftlich geschriebene Text 17, der eher zum Funktionalstil der Alltagssprache tendiert, bildet hiervon eine Ausnahme.

Die konventionalisierten Zwischenüberschriften der einzelnen Gliederungspunkte sowie die in Publikationen häufig zu findenden Abschnitte *abstract* und *acknowledgements* wurden nicht als integrale Textbestandteile angesehen und daher bei der Zählung nicht berücksichtigt.

Die Ergebnisse:

	Text 17		Text 18		Text 19	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	1.039	1.037.87	325	321.29	557	552.19
2	387	386.54	117	117.05	135	135.57
3	187	191.71	81	83.39	105	116.14
4	89	83.65	43	49.18	76	69.77
5	26	28.34	33	22.03	42	31.44
6	9	7.71	3	7.90	4	11.34
7	1	2.18	2	3.16	2	4.55
	$a = 1.3608$; $b = 0.1390$; $\alpha = 0.4449$. $X_3^2 = 1.49$; $P = 0.68$.		$a = 1.7937$; $b = 0.0940$; $\alpha = 0.5086$; $X_1^2 = 1.62$; $P = 0.20$.		$a = 1.8030$; $b = 0.0141$; $\alpha = 0.4706$; $X_1^2 = 1.68$; $P = 0.20$.	

	Text 20		Text 21	
x	n_x	NP_x	n_x	NP_x
1	716	716.87	448	444.31
2	262	263.49	230	230.22
3	185	181.76	112	109.90
4	89	86.05	46	58.84
5	27	30.56	36	28.23
6	9	8.68	16	16.50
7	0	2.05	-	-
8	0	0.41	-	-
9	1	0.07	-	-
10	1	0.06	-	-
	$a = 1.4208$; $b = 0.0144$; $\alpha = 0.5778$; $X_3^2 = 0.72$; $P = 0.87$.		$a = 2.0123$; $b = 0.3667$; $\alpha = 0.3444$; $X_2^2 = 5.02$; $P = 0.08$.	

Text 17: John H. Lawton: Eat Caribbean bananas.
Anmerkungen: „Caribbean" wurde nach der britischen Aussprache als viersilbig gewertet. „doesn't" wurde als ein zweisilbiges Wort angesehen. „different" wurde als dreisilbig gezählt.

Text 18: M. Schultz: Microscopic investigations on tumorous lesions from Christian Salaya (Egyptian Nubia).
Anmerkungen: „Ind. K-68/2" wurde *Index K-sixtyeight slash two* gelesen, hat also vier Wörter à zwei, vier und zweimal einer Silbe. „interesting" wurde als dreisilbig gewertet.

Text 19: Kazuo Kawada: Production of Sexuals in the Aphid Acyrthosiphon kondoi in Japan (Homoptera: Aphidae).
Anmerkungen: „kondoi" wurde als zweisilbig gewertet, die japanischen Ortsbezeichnungen „Kyushu" und „Hokkaido" als drei- bzw. viersilbig. „N' part" wurde als *north part* und „1986-06-20" als *twentieth of June nineteeneightysix* gelesen. „43° 03' N" und entsprechende geographische Angaben wurde *fourtythree degrees, three minutes north* aufgelöst. „16L:8D" wurde als *sixteen L to eight D* ausgewertet, „L_4" als *L four*. „Sx-♀♀" ist die Abkürzung für *sexualis-females*, „Sx-♂♂" entsprechend für *sexualis-males*. „temperate" wurde als dreisilbig gezählt. „(Harris 1776)", „(Linnaeus 1758)" und „(Sulzer 1775)" wurden jeweils als zwei Wörter gewertet. „highest %" wurde als *highest percentage* gelesen.

Text 20: William Mahaney, R.G.V. Hancock and M. Inoue: Geochemistry and Clay Mineralogy of Soils Eaten by Japanese Macaques.
Anmerkungen: „NaCl" wurde als *sodiumchlorite* gelesen, „metahalloysite" wurde als Sechssilber gezählt, „Inoue" als Dreisilber. „Ah", „Bw", „Cox" und „Cu" sind Abkürzungen für Bodenhorizonte und wurden buchstabenweise gelesen. „H_2O" wurde als *water* ausgewertet. „10YR" wurde als dreisilbig und „Al_2OH_6" als sechssilbig gezählt.

Text 21: Audun Slettan, I. Olsaker and O. Lie: Isolation and characterization of variable $(GT)_n$ repetitive sequences from Atlantic salmon. Salmo salar L.
Anmerkungen: „$(GT)_n$ wurde als ein Wort mit drei Silben gewertet, „$(dG\text{-}dT)_n$" als ein Wort mit sechs Silben. „Sau3A" wurde als *Sau-three-A*, „SK+" als *S-K-plus* gelesen. „kb" wurde als *kilobases* ausgewertet, „10^4" als *ten to the power four*.

2. Zusammenfassung

Es kann festgestellt werden, daß sich alle Texte mit der gemischten Poisson-Verteilung modellieren lassen, obwohl sie sehr unterschiedlichen Textsorten angehören, unter entsprechend verschiedenen Bedingungen entstanden sind bzw. jeweils anderen Anforderungen genügen müssen und in ihrer Entstehungszeit weit auseinanderliegen. Dieses Ergebnis stimmt sehr gut mit der Untersuchung von Riedemann (1994) überein, bei der 57 von 60 Pressetexten der gemischten Poisson-Verteilung entsprechen. In drei Fällen konnte Riedemann außerdem die Hirata-Poisson-Verteilung anpassen, die sich bisher bei französischen Texten als geeignet erwies.[6]

Literatur:

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Collins English Dictionary ^{3}(1991). Hg. v. J.M. Sinclair. Sydney: Harper Collins Publishers.

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In P. Schmidt (Hg.), *Glottometrika 15* (S. 158-165), Trier: WVT.

[6] Dieckmann & Judt (1996); Feldt, Janssen & Kuleisa (1997).

Feldt, S., Janssen, M., & Kuleisa, S. (1997). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In diesem Band.

Fitzgerald, F., Scott (1993). *The Letters of F. Scott Fitzgerald.* Ed. by A.Turnbull, London: Budley.

Frost, R. (1964). *Selected Letters by Robert Frost.* Ed. by L. Thompson, London: Holt.

Frost, R. (1963). *The Letters of Robert Frost to Louis Untermeyer.* New York: Holt.

Kawada, K. (1992). Production of Sexuals in the Aphid Acyrthosiphon kondoi in Japan (Homoptera: Aphidinea: Aphidae). *Entomologia Generalis. Journal of General and Applied Entomology.* 17/2, 115-119.

Lawton, J. H. (1993). Eat Caribbean bananas. *Oikos. A Journal of Ecology*, 66/1, 3-4.

Mahaney, W., Hancock, R.G.V., & Inoue, M. (1993). Geochemistry and Clay Mineralogy of Soils Eaten by Japanese Macaques. *Primates. A Journal of Primatology*, 34/1, 85-91.

Orwell, G. (1970). *The Collected Essays, Journalism & Letters of George Orwell, Vol. I (1920-40).* Ed. by S. Orwell and J. Angus. London: Penguin.

Riedemann, H. (1994). Wortlängen in der Sprache der deutschen und englischen Tages- und Wochenpresse. Staatsexamensarbeit. Göttingen.

Schultz, M. (1993). Microscopic investigations on tumorous lesions from Christian Salaya (Egyptian Nubia). *Anthropologischer Anzeiger. Offizielles Publikationsorgan der Schweizerischen Gesellschaft für Anthropologie (Société Suisse d'Anthropologie)*, 51/2, 117-121.

Slettan, A., Olsaker I., & Lie, O. (1993). Isolation and characterization of variable (GT), repetitive sequences from Atlantic salmon, Salmo salar L. *Animal Genetics, Immunogenetics. Biochemical Genetics and Molecular Genetics*, 24/3, 195-197.

Webster's Ninth New Collegiate Dictionary (1987). Chief ed. C. Frederick. Springfield, Mass.: Merriam Webster.

Williams, W.C. (1957). *The Selected Letters of William Carlos Williams.* Ed. by J. C. Thirwall. New York: McDowell.

Glottometrika 16, 1997, 108-120

Zur Wortlängenhäufigkeit in griechischen Koine-Texten

Jannetje Egbers, Claudia Groen, Ralf Podehl, Esther Rauhaus

1. In dieser Untersuchung geht es darum, die Häufigkeit des Auftretens von Wörtern verschiedener Länge in griechischen Koine-Texten festzustellen und zu überprüfen, ob eines der zu diesem Zweck entwickelten Modelle an die Daten dieser Texte angepaßt werden kann (vgl. dazu z.B. Wimmer & Altmann, 1996).

2. Aus dem Bereich der griechischen Koine wurden zwei unterschiedliche Textgruppen bearbeitet:

a) Texte des Neuen Testaments[1]
b) Texte von Epiktet

Das Neue Testament, das in dem Zeitraum von 50 bis 150 n.Chr. verfaßt wurde, stellt auch bezüglich der Verfasserschaft keine homogene Einheit dar. Es wurde in der damaligen griechischen Volkssprache, der Koine (ἡ κοινὴ διάλεκτος) geschrieben. Die Koine meint das Griechisch in der Zeit des Hellenismus, dem Zeitalter, das von der Begegnung der antiken Welt mit dem Orient geprägt war. Sie stellt das Übergangsstadium vom klassischen Griechisch zum späteren Mittelgriechisch und dem gesprochenen Neugriechisch dar.

Den ersten Anstoß für die Bildung der Koine gab die Gründung des ersten attischen Seebundes 478/77 (vgl. Wilsdorf, 1986:539) Aufgrund der Vormachtstellung Athens war die sich entwickelnde Koine stark vom Attischen geprägt, sie nahm aber auch ionische Elemente in sich auf. Als Reichssprache des Makedonenreiches unter Alexander dem Großen und den Diadochen verdrängte sie die ursprünglichen griechischen Dialekte. Für die Ausbildung der hellenistischen Koine in römischer Zeit war entscheidend, daß die Römer Griechisch als Amtssprache in den Provinzen im Osten wählten. Griechisch war die Voraussetzung für Verkehr und Beruf und weniger ein Zeichen besonderer Bildung (Rehkopf, 1986:229).

[1] Da die neutestamentlichen Texte nicht mehr im Original vorliegen, dient als Grundlage die von Nestle und Aland rekonstruierte Fassung.

In Palästina hatte sich bereits in vorrömischer Zeit der Hellenismus in Kultur und Sprache durchgesetzt. Im 1. Jahrhundert n. Chr. war in Palästina Griechisch neben dem Aramäischen die gesprochene Sprache.

Obwohl der Stoff des Neuen Testaments oft in einer anderen Sprache überliefert wurde (z.B. aramäisch), ist der griechische Text des Neuen Testaments keine Übersetzung aus dieser anderen Sprache, sondern ist in griechischer Sprache verfaßt. Im Vergleich zur Koine weisen neutestamentliche griechische Texte wiederum charakteristische Merkmale auf, da sie stark durch die Sprache des überlieferten Stoffes beeinflußt sind und meistens von Verfassern stammen, deren Muttersprache nicht das Griechische war (vgl. Wilkenhauser & Schmidt, 1973:190).

Die Verbreitung des Griechischen als Allgemeinsprache führte gegenüber dem klassischen Griechisch zu einer Vereinfachung in Deklination, Konjugation und Syntax. Ebenso kam es zu einer Veränderung im Wortschatz. Wörter bestimmter Bedeutungskategorien wurden häufiger benutzt oder in ihrer ursprünglichen Bedeutung abgewandelt; ebenso sind im Neuen Testament Neologismen vorhanden. Einflüsse anderer Sprachen auf die Koine sind relativ gering geblieben. Ins neutestamentliche Griechisch sind Semitismen und Latinismen sowie einige Lehnwörter aus dem Persischen und Koptischen eingegangen (vgl. Rehkopf, 1986:231).

Allgemein läßt sich über den Stil der Sprache im Neuen Testament sagen, daß sie weder der vornehmen attischen Literatursprache noch der Umgangssprache zuzurechnen ist. Das neutestamentliche Griechisch „...ist als nichtattizierende, überwiegend nicht-literarische Koine sowohl mit den literarischen wie unliterarischen Papyri als auch der zeitgenössischen Fachprosa und mit den Schriftstellern wie etwa Epiktet.... zu vergleichen“ (Blass & Debrunner, 1986:§3).

Der stoische Philosoph Epiktet, ein griechischer Muttersprachler, lebte von ca. 50 bis 120 n. Chr. und verwendete eine ähnliche Variante der Koine wie die Verfasser der neutestamentlichen Texte. Seine Diatribe sind allerdings nicht von ihm selbst verfaßt worden, sondern von einem seiner Schüler, Flavius Arrianus, der die Gespräche mit seinem Meister zur eigenen Erinnerung ziemlich wörtlich aufgeschrieben hat. Nicht weiter überarbeitet oder stilisiert (vgl. Schwartz, 1986: Sp. 1232) gab Arrian später diese Niederschriften an Freunde weiter; sie sind zur Basis der Textüberlieferung geworden. Da diese nur auf eine einzige archetypische Handschrift aus dem 11. Jahrhundert zurückgeht, gibt es keine voneinander abweichenden Lesarten, die womöglich auch eine unterschiedliche Wortlängenverteilung aufweisen würden (vgl. Oldfather, 1926:xxxiff).
Der griechische Text stellt einzelne Unterrichtseinheiten Epiktets dar, die jeweils ein abgeschlossenes philosophisches Thema behandeln. Die Gliederung in Sinnabschnitte wurde von Arrian übernommen.

Durch die Zugehörigkeit des Werkes zur Gattung der Diatribe sind bestimmte sprachliche Ausdrücke vorgegeben. Diese populärwissenschaftlichen Vorträge sind von einem sehr lockeren, einfachen Gesprächston bestimmt. Der dialogische Stil ist schlicht, die Sätze sind extrem kurz; oft handelt es sich um Ellipsen, um einfache Ausrufe oder kurze Fragen. Durch den umgangssprachlichen Grundton und das Verfassen der Texte hauptsächlich im Präsens kann man erwarten, daß lange Wörter eher selten auftauchen, da die Flexionsendungen im Präsens eher kurz sind.

3. Alle ausgewerteten Texteinheiten wurden vollständig daraufhin untersucht, wie oft Wörter verschiedener Silbenzahl in ihnen vorkommen. „Wort" wurde als orthographisches Wort (vgl. Bünting & Bergenholtz, 1986:39ff) bestimmt; die Länge der Wörter wurde nach der Zahl der in ihnen enthaltenen Silben angegeben. Kriterium für die Silbenzahl ist die Zahl der Vokale oder Diphthonge im Wort. Das Verfahren stimmt damit mit dem in Best & Zhu (1994) verwendeten überein.

Die griechische Buchstabenfolge <υι> wurde nur dann als Diphthong gelesen, wenn ein weiterer Vokal folgte. Im Griechischen gibt es keine Halbvokale. In Lautverbindungen, in denen <ι> vor Vokal vorkommt, wurde das <ι> als vokalisch gewertet (z.B. κύριος - dreisilbig).

Ein Problemfall bei der Bestimmung der Silbenanzahl der griechischen Sprache bilden die durch Elision entstandenen nullsilbigen Wörter. Solche Enklitika wurden in Verbindung mit dem folgenden Wort als *ein* Wort gezählt. Ein besonders häufig auftretender Problemfall dieser Art ist die Elision des <ε> bei <δε>. Die semantische Bedeutungslosigkeit dieser Partikel im Koine-Griechischen rechtfertigt diese Zählung.

Die griechischen Verschriftungsversuche lateinischer Eigennamen wie „Vespasian" (Ὀυσπασιάνος) und „Helvidius" (Ελουίδιος) (vgl. Epiktet, Diatribe, Buch I, Kapitel 2), die den lateinischen Halbvokal <u> bzw. <v> als Diphthong wiedergegeben, wurden konsonantisch gezählt.

4. Die erarbeiteten Daten wurden mit dem Altmann-Fitter (1994) daraufhin untersucht, welchem mathematischen Modell sie folgen. Es zeigte sich, daß an alle Texte die Hyperpoisson-Verteilung in der 1-verschobenen Form angepaßt werden konnte; für einzelne Texte mußte diese aber modifiziert werden (siehe unten). Die Formel für die Hyperpoisson-Verteilung lautet:

$$P_x = \frac{a^x}{{}_1F_1\,(1; b; a)\, b^{(x)}}, \quad x = 0, 1, 2, \ldots$$

Dabei ist ${}_1F_1\,(1;\, b;\, a)$ die konfluente hypergeometrische Funktion,

$${}_1F_1\,(1;\, b;\, a) \;=\; \sum_{j=0}^{\infty} \frac{a^j}{b^{(j)}}$$

und

$$b^{(j)} = b\,(b+1)\,(b+2)\,\ldots\,(b+j-1).$$

Die Anpassungsgüte wurde mit dem Chiquadrat-Test überprüft. Eine Anpassung eines Modells an die Daten betrachten wir dann als zufriedenstellend, wenn die Wahrscheinlichkeit für das berechnete oder noch größere X^2, $P(X^2) \geq 0{,}05$ ist oder, besonders bei etwas längeren Texten, $C \leq 0{,}02$, wobei $C = X^2/N$ ist. a und b sind Parameter, x ist die Wortlänge, n_x die Zahl der Wörter im Text mit der entsprechend Silbenzahl x, NP_x der theoretische Wert dazu.] bedeutet die Zusammenfassung von Klassen.

Die Ergebnisse stellen sich wie folgt dar:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	71	70.33	35	33.65	56	55.99
2	89	82.57	19	21.93	53	52.99
3	39	51.63	15	13.14	33	31.00
4	29	22.00	12	7.29]	10	13.13
5	7	7.11	2	7.00]	6	4.36
6	1	2.36	-	-	1	1.55
	a = 1.338; b = 1.40; X^2_3 = 6.60; P = 0.09.		a= 7.471; b= 11.466; X^2_1 = 0.71; P = 0.40.		a= 1.532; b= 1.619; X^2_3= 1.68; P= 0.64.	

Text 1: Markusevangelium 6,30-44.
Text 2: Markusevangelium 1,7-11.
Text 3: Matthäusevangelium 14,13-21.

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	61	60.68	52	52.86	66	64.39
2	53	49.16	60	51.22	42	45.76
3	21	27.91	22	33.59	28	27.55
4	15	12.20	19	16.65	17	14.39
5	5	4.33	9	6.63	8	6.64
6	1	1.74	2	3.07	1	2.74
7	-	-	-	-	1	1.55
	$a = 1.897$; $b = 2.341$; $X_3^2 = 3.07$; $P = 0.38$.		$a = 2.030$; $b = 2.094$; $X_3^2 = 7.05$; $P = 0.07$.		$a = 3.942$; $b = 5.546$; $X_4^2 = 2.40$; $P = 0.66$.	

Text 4: Matthäusevangelium 3,11-17.
Text 5: Lukasevangelium 9,10-17.
Text 6: Lukasevangelium 3,15-22.

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	68	66.19	68	67.09	152	149.31
2	65	63.25	68	66.68	125	120.87
3	36	39.86	31	34.78	55	65.80
4	18	18.74	14	12.30	31	26.99
5	10	7.02	4	4.17	12	12.06
6	1	2.97	-	-	-	-
	$a = 1.851$; $b = 1.937$; $X_3^2 = 3.05$; $P = 0.38$.		$a = 1.098$; $b = 1.105$; $X_2^2 = 0.69$; $P = 0.71$.		$a = 1.663$; $b = 2.054$; $X_2^2 = 2.56$; $P = 0.28$.	

Text 7: Johannesevangelium 6,1-13.
Text 8: Johannesevangelium 1,24-34.
Text 9: Johannesevangelium 2,1-22.

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	287	289.29	128	125.71	190	182.15
2	286	263.01	112	110.00	210	200.22
3	134	160.85	53	60.26	126	145.62
4	74	74.12	32	24.03	76	79.13
5	37	27.39	5	10.02	48	34.32
6	6	8.45	-	-	9	17.59
7	2	2.91	-	-	-	-
	$a = 1.868$; $b = 2.055$; $X_4^2 = 10.86$; $P = 0.03$; $C = 0.01$.		$a = 1.465$; $b = 1.674$; $X_1^2 = 1.21$; $P = 0.27$.		$a = 2.149$; $b = 1.955$; X_2^2 4.09; $P = 0.13$.	

Text 10: Zweiter Brief des Paulus an die Thessalonicher.
Text 11: Brief des Paulus an Philemon.
Text 12: Brief des Paulus an Titus.

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	483	474.47	436	420.68	128	128.65
2	491	509.07	325	344.41	152	140.64
3	358	366.45	218	226.84	89	101.30
4	224	198.48	146	124.97	51	54.41
5	89	86.16	76	59.17	31	23.30
6	28	31.21	12	24.56	6	8.30
7	5	9.70	0	9.08	3	3.43
8	1	3.47	1	4.31	-	-
	$a = 2.187$; $b = 2.039$; $X_5^2 = 8.71$; $P = 0.12$.		$a = 3.369$; $b = 4.114$; $X_2^2 = 6.21$; $P = 0.045$; $C = 0.005$.		$a = 2.112$; $b = 1.932$; $X_4^2 = 5.86$; $P = 0.21$.	

Text 13: Erster Brief des Paulus an Timotheus.
Text 14: Zweiter Brief des Paulus an Timotheus.
Text 15: Der Brief des Judas.

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	93	93.35	75	70.47	242	237.62
2	91	84.47	66	62.02	216	226.49
3	38	44.51	31	42.24	115	104.91
4	15	16.55	36	23.47	36	32.09
5	8	6.14	8	17.81	0	7.33
6	-	-	-	-	1	1.59
	$a = 1.262$; $b = 1.394$; $X_2^2 = 2.18$; $P = 0.34$.		$a = 3.015$; $b = 3.426$; $X_1^2 = 3.72$; $P = 0.05$.		$a = 0.901$; $b = 0.945$; $X_1^2 = 1.93$; $P = 0.17$.	

Text 16: Der zweite Brief des Johannes.
Text 17: Der dritte Brief des Johannes.
Text 18: Johannesevangelium Kapitel 20.

	Text 19		Text 20		Text 21	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	68	64.48	196	195.23	315	308.58
2	52	48.80	208	202.50	200	207.80
3	21	29.82	94	102.88	109	117.36
4	22	15.28	36	34.61	74	57.07
5	6	10.66	12	10.81	31	38.23
	$a = 3.172$; $b = 4.191$; $X_1^2 = 3.18$; $P = 0.07$.		$a = 0.996$; $b = 0.960$; $X_2^2 = 1.11$; $P = 0.57$.		$a = 3.500$; $b = 5.198$; $X_2^2 = 7.40$; $P = 0.03$; $C = 0.01$.	

Text 19: Johannesevangelium 7,53-8,11.
Text 20: Johannesevangelium Kapitel 21.
Text 21: Diatribe Buch I, Kapitel 1.

	Text 22		Text 23		Text 24	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	315	310.39	315	310.98	140	138.56
2	265	267.22	239	245.74	111	104.44
3	144	158.25	159	146.27	50	63.47
4	92	071.43	54	69.83	37	32.31
5	22	026.04	36	27.83	16	14.15
6	5	7.97	11	13.38	7	8.12
7	0	2.10	-	-	-	-
8	1	0.65	-	-	-	-
	$a = 1.898$; $b = 2.204$; $X_4^2 = 10.10$; $P = 0.04$; $C = 0.01$.		$a = 2.412$; $b = 3.052$; $X_3^2 = 7.75$; $P = 0.05$.		$a = 3.136$; $b = 4.160$; $X_3^2 = 4.36$; $P = 0.23$.	

Text 22: Diatribe Buch I, Kapitel 2.
Text 23: Diatribe Buch I, Kapitel 4.
Text 24: Diatribe Buch I, Kapitel 8.

	Text 25		Text 26		Text 27	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	358	356.92	338	332.67	248	247.28
2	317	302.09	256	269.19	211	199.63
3	163	184.46	159	155.90	100	118.64
4	89	88.09	80	70.31	60	55.78
5	44	34.54	27	25.96	27	21.70
6	7	11.49	5	11.00	5	7.20
7	4	4.42	-	-	2	2.81
	$a = 2.192$; $b = 2.590$; $X_4^2 = 7.63$; $P = 0.11$.		$a = 2.037$; $b = 2.518$; $X_3^2 = 5.43$; $P = 0.14$.		$a = 2.252$; $b = 2.789$; $X_4^2 = 6.08$; $P = 0.19$.	

Text 25: Diatribe Buch I, Kapitel 9.
Text 26: Diatribe Buch I, Kapitel 12.
Text 27: Diatribe Buch I, Kapitel 17.

	Text 28		Text 29		Text 30	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	278	274.35	259	257.47	260	256.43
2	213	205.35	225	223.67	165	173.57
3	92	113.35	122	127.94	93	91.59
4	63	49.55	57	54.55	47	39.60
5	20	17.93	23	18.54	13	14.50
6	1	5.54	3	6.85	3	4.61
7	1	1.96	-	-	1	1.73
	$a = 2.102$; $b = 2.808$; $X_2^2 = 8.46$; $P = 0.02$; $C = 0.01$.		$a = 1.675$; $b = 1.928$; $X_3^2 = 3.63$; $P = 0.30$.		$a = 2.394$; $b = 3.536$; $X_4^2 = 2.89$; $P = 0.58$.	

Text 28: Diatribe Buch I, Kapitel 18.
Text 29: Diatribe Buch I, Kapitel 19.
Text 30: Diatribe Buch I, Kapitel 22.

	Text 31		Text 32		Text 33	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	240	237.82	358	355.66	271	263.79
2	188	180.29	263	252.14	197	207.90
3	83	102.56	117	137.78	119	123.05
4	61	46.69	67	61.24	71	58.31
5	14	17.71	27	22.94	23	23.04
6	7	7.96	8	10.26	5	7.81
7	-	-	-	-	1	3.15
	$a = 2.279$; $b = 3.006$; $X_3^2 = 9.36$; $P = 0.03$; $C = 0.02$.		$a = 2.383$; $b = 3.362$; $X_3^2 = 5.36$; $P = 0.15$.		$a = 2.377$; $b = 3.016$; $X_4^2 = 6.10$; $P = 0.19$.	

Text 31: Diatribe Buch II, Kapitel 2.
Text 32: Diatribe Buch II, Kapitel 5.
Text 33: Diatribe Buch II, Kapitel 6.

	Text 34		Text 35	
x	n_x	NP_x	n_x	NP_x
1	200	192.95	241	237.62
2	151	159.77	204	200.27
3	101	102.49	102	116.44
4	59	53.66	60	51.68
5	34	23.73	21	18.54
6	0	9.08	4	7.47
7	1	4.35	-	-
	$a = 2.847$; $b = 3.438$; $X_2^2 = 1.42$; $P = 0.49$.		$a = 1.875$; $b = 2.224$; $X_3^2 = 5.17$; $P = 0.16$.	

Text 34: Diatribe Buch II, Kapitel 9.
Text 35: Diatribe Buch II, Kapitel 11.

Bei einigen Daten zeigt sich eine konsequente Verschiebung einiger Häufigkeiten von $x = 3$ auf $x = 4$ (Anomalien sieht man sehr gut in den Texten: 36 und 38). Dies bedeutet, daß im Griechischen dieser Zeit eine langsame Verschiebung zu einem anderen Attraktor stattfand. Vorläufig lassen sich die Anomalien mit einer Modifikation des Modells erfassen. Wir behalten die 1-verschobene Hyperpoisson-Verteilung bei und modifizieren sie folgendermaßen:

$$P'_x = \begin{cases} P_x & x = 1, 2, 5, 6, \ldots \\ P_3 (1 - \alpha) & x = 3 \\ P_4 + \alpha P_3 & x = 4 \end{cases}$$

wobei P_x die ursprüngliche 1-verschobene Hyperpoisson-Verteilung darstellt. Eine derartige Modifikation hat sich bereits im Tschechischen und im Türkischen ergeben (s. Uhlířová, 1995; Altmann, Erat & Hřebíček, 1996).

Die Resultate sind wie folgt:

	Text 22		Text 36		Text 37	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	315	316.24	309	312.90	237	234.37
2	265	275.98	212	205.46	223	220.53
3	144	127.78	86	84.43	98	100.64
4	92	89.05	97	89.39	73	69.81
5	22	21.87	32	31.04	15	15.85
6	5	6.04	19	13.84	4	4.08
7	0	1.43	2	9.07	1	1.09
8	1	0.36	-	-	-	-
	$a = 1.617$; $b = 1.853$; $\alpha = 0.1831$; $X_3^2 = 3.12$; $P = 0.37$.		$a = 5.550$; $b = 8.453$; $\alpha = 0.300$; $X_2^2 = 1.12$; $P = 0.57$.		$a = 1.417$; $b = 1.506$; $\alpha = 0.193$; $X_3^2 = 0.33$; $P = 0.95$.	

Text 22: Diatribe Buch I, Kapitel 2.
Text 36: Diatribe Buch I, Kapitel 7.
Text 37: Diatribe Buch I, Kapitel 14.

	Text 28		Text 38		Text 31	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	278	282.86	168	163.79	240	243.83
2	213	214.58	178	174.48	188	185.97
3	92	83.67	67	70.33	83	73.99
4	63	63.70	69	60.81	61	62.00
5	20	13.59	10	11.92	14	14.53
6	1	3.57	2	3.60	7	5.56
7	1	0.81	-	-	-	-
	$a = 1.607$; $b = 2.114$; $\alpha = 0.244$; $X_3^2 = 4.47$; $P = 0.11$.		$a = 1.255$; $b = 1.177$; $\alpha = 0.300$; $X_2^2 = 2.46$; $P = 0.29$.		$a = 1.871$; $b = 2.454$; $\alpha = 0.266$; $X_2^2 = 1.59$; $P = 0.45$.	

Text 28: Diatribe Buch I, Kapitel 18.
Text 38: Diatribe Buch I, Kapitel 24.
Text 31: Diatribe Buch II, Kapitel 2.

	Text 39		Text 40	
x	n_x	NP_x	n_x	NP_x
1	260	258.00	105	105.38
2	227	225.26	87	80.82
3	111	110.86	27	32.01
4	85	86.23	29	24.29
5	31	24.67	3	5.27
6	6	7.93	1	1.80
7	1	2.88	-	-
	$a = 2.034$; $b = 2.330$; $\alpha = 0.195$; $X_3^2 = 3.37$; $P = 0.34$.		$a = 1.625$; $b = 2.118$; $\alpha = 0.240$; $X_2^2 = 3.51$; $P = 0.17$.	

Text 39: Diatribe Buch II, Kapitel 10.
Text 40: Johannesevangelium 1-1,19.

5. Die Untersuchung hat folgendes ergeben:
Alle bearbeiteten Texte folgen der Hyperpoisson-Verteilung in der 1-verschobenen Form; für einige Texte mußten Modifikationen eingesetzt werden. Hier deutet sich ein Strukturwandel des Griechischen in der Wortlängenverteilung an. Dennoch machen die untersuchten Texte einen unerwartet homogenen Eindruck, zumal auch die gleiche Modifikation in beiderlei Texten anwendbar ist. Die Untersuchung darf nicht als repräsentativ für das Griechische insgesamt gewertet werden. In anderen Arbeiten hat sich gezeigt, daß verschiedene Textsorten und Zeitabschnitte einer Sprache teilweise unterschiedlichen Modellen folgen (vgl. z.B. zum Deutschen Laass, 1996). Ob dies auch für älteres oder auch für jüngeres Griechisch gilt, muß weiter untersucht werden.

Literatur

Altmann, G., Erat, E., & Hřebíček, L. (1996). Word Length Distribution in Turkish Texts. In P. Schmidt (Hg.), *Glottometrika 15* (S. 195-204), Trier: WVT.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk, (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Blass, F., & Debrunner, A. (1984). *Grammatik des neutestamentlichen Griechisch.* Bearbeitet von F. Rehkopf. 16. Auflage. Göttingen: Vandenhoeck & Ruprecht.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax.* Frankfurt: Athenäum.

Laass, F. (1996). Zur Häufigkeit der Wortlängen in deutschen Lesebuchtexten. In P. Schmidt (Hg.), *Glottometrika 15* (S.181-194), Trier: WVT.

Nestle, K., Aland, E., & Aland, K. (1986). *Das neue Testament. Griechisch und Deutsch.* 26. Auflage. Stuttgart: Deutsche Bibelgesellschaft.

Oldfather, W.A. (Hg. und Übers.) (1926). *Epictetus. The Discourses as reported by Arrian, the Manual, Fragments.* 2 Bände. In Capps, Page, Rose (Hg.), The Loeb Classical Library, London: William Heinemann; New York: G.P. Putnam's sons.

Rehkopf, F. (1986). Griechisch des Neuen Testaments. In G. Müller (Hg.), *Theologische Realenzyklopädie* (S. 228-235), Berlin u.a.: De Gruyter.

Schwartz, E. (1896). Arrianus. In G. Wissowa (Hg.), *Paulys Realenzyklopädie der classischen Altertumswissenschaften.* 2.Band (Sp. 1228 - 1247), Stuttgart: J.B. Metzlerscher Verlag.

Uhlířová, L. (1995). O jednom modelu rozložení délky slov. Mskr.

Wilkenhauser, A., & Schmid, J. (1973). *Einleitung in das Neue Testament.* 6. Auflage, Freiburg u.a.: Herder.

Wilsdorf, H. (1986). Sprachkenntnisse: In J. Irmscher (Hg.), *Lexikon der Antike* (S. 539), Leipzig: VEB Bibliographisches Institut.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In P.Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

SOFTWARE

Altmann-FITTER (1994), Lüdenscheid: RAM-Verlag.

Glottometrika 16, 1997, 121-126

Wortlängenhäufigkeit in Plinius - Briefen

Winfred Röttger, Anja Schweers

1. Die vorliegende Untersuchung steht im Zusammenhang mit dem Göttinger Wortlängenprojekt (Best & Zhu, 1994) und ermittelt die Häufigkeit des Vorkommens von Wörtern unterschiedlicher Länge in abgeschlossenen Texten. Grund dafür ist die Annahme, daß Wörter unterschiedlicher Länge in Texten nicht chaotisch, sondern stochastischen Gesetzen folgend auftreten. Die Kriterien für „Wort" und Silbe sind die gleichen wie in Best & Zhu (1994:20): Als Wort gilt das orthographische Wort (Bünting & Bergenholtz, 1989:36f); die Zahl der Silben im Wort wird danach bestimmt, wieviele Vokale im Wort enthalten sind.

2. Am Beispiel der Briefe des Plinius behandeln wir zum erstenmal lateinische Texte. Die Briefe des Plinius sind für diese Untersuchung besonders geeignet, weil sie Kriterien erfüllen, die sich für eine Modellierbarkeit als günstig erwiesen haben: denn die Texte dieses Funktionalstils bieten ein - für die Untersuchung wünschenswertes - sprachlich homogenes Bild insofern, als sie von Plinius selbst geschrieben wurden und nicht mehr als 2000 Wörter, häufig auch weniger, aufweisen. Die Auswahl der einzelnen Briefe aus dem umfangreichen Textkorpus (247 Briefe ohne die Trajankorrespondenz) orientierte sich an einer durchschnittlichen Wortzahl von ca. 100 bis 1000 Wörtern, erfolgte ansonsten jedoch willkürlich. Texte mit längeren Zitaten wurden allerdings von vornherein ausgeschlossen.

Zur Datenaufnahme: Griechische Wörter und Grußformeln wurden nicht ausgewertet. Abkürzungen (vgl. Brief 6, 16: *kal.*) und Zahlsymbole wurden als ausgeschrieben betrachtet, so z. B. *D* als *quingenti*. Eigennamen wurden mit ausgezählt. Enklitika (*-que,-ve,-ne*) wurden als Teil eines Wortes behandelt. Die Auswertung orientierte sich an der Aussprache, wie sie in den Standardgrammatiken des Lateinischen beschrieben wird (z. B. Rubenbauer, Hofmann & Heine, 1977:5ff).

Die Briefe des Plinius sind in drei Handschriftengruppen überliefert. Der vorliegenden Untersuchung liegt die anerkannte textkritische Ausgabe von R. A. B. Mynors (1982) zugrunde.

3. Die Wortlängen in allen Texten, bis auf Brief 8,14 und 8,24, folgen der positiven Binomialverteilung

$$P_x = \binom{n}{x} \frac{p^x q^{n-x}}{1-q^n}, \quad x = 1, 2, \ldots n.$$

In zahlreichen Fällen, besonders aber in Briefen 8,14 und 8,24, kann man als Alternative die 1-verschobene Palm-Poisson-Verteilung

$$P_x = \frac{R_{(x-1)}\, a^{x-1}}{T}, \quad x = 1,2, \ldots, R+1$$

verwenden, mit T als Normierungskonstante, die wie in Wimmer et al. (1994:102) gezeigt aus dem gleichen allgemeinen Ansatz wie die Binomialverteilung ableitbar ist.

Die Anpassung der Daten an das Modell betrachten wir als zufriedenstellend, wenn $P \geq 0.05$ oder, besonders bei längeren Texten, $C \leq 0.02$. Diese Bedingungen sind für sämtliche untersuchten Briefe erfüllt. In den folgenden Tabellen bedeutet: x die Wortlängenklasse, n_x die beobachtete Häufigkeit, NP_x die nach der positiven Binomialverteilung und NP_{x*} die nach der Palm - Poisson - Verteilung berechneten Werte. P ist die Überschreitungswahrscheinlichkeit des Chiquadrats, C der Diskrepanzkoeffizient $C = X^2/N$. Die übrigen Größen sind Parameter.

4. Die Untersuchung erbrachte im einzelnen folgende Ergebnisse:

	Text 1: Brief 2,17		Text 2: Brief 3,9		Text 3: Brief 8,14		
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	NP_x*
1	228	215.37	297	279.53	325	300.63	314.59
2	328	350.39	352	378.06	296	359.87	320.05
3	311	304.02	300	306.79	276	266.67	260.49
4	150	148.38	187	165.97	193	136.81	159.01
5	40	38.62	58	62.85	37	51.47	64.71
6	4	4.22	18	17.00	5	14.67	13.17
7	-	-	2	3.80	2	3.88	-
	$n = 6$; $p = 0.3942$; $X_3^2 = 2.41$; $P = 0.49$.		$n = 11$; $p = 0.2129$; $X_4^2 = 6.97$; $P = 0.14$.		NP_x: $n = 15$; $p = 0.1460$; $X_1^2 = 18.04$; $P = 0.00$; $C = 0.0159$.		NP_{x*}: $a = 0.2035$; $R = 6$; $X_1^2 = 3.09$; $P = 0.08$.

	Text 4: Brief 1,19		Text 5: Brief 2,16		Text 6: Brief 3,12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	31	29.82	38	37.29	28	28.96
2	38	35.10	39	38.63	34	31.76
3	22	26.01	23	25.41	20	22.00
4	11	13.60	13	11.91	12	10.79
5	10	7.47	6	5.76	5	5.49
	$n = 19$; $p = 0.1157$; $X_2^2 = 2.27$; $P = 0.32$.		$n = 22$; $p = 0.0898$; $X_2^2 = 0.36$; $P = 0.84$.		$n = 20$; $p = 0.105$; $X_2^2 = 0.55$; $P = 0.76$.	

	Text 7: Brief 5,21		Text 8: Brief 6,1		Text 9: Brief 7,3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	44	45.66	28	27.93	39	37.91
2	55	49.95	30	29.17	46	46.49
3	31	34.61	16	17.76	31	33.79
4	16	17.04	8	6.95	19	16.11
5	8	6.34	1	1.81	5	5.27
6	2	2.40	1	0.38	1	1.43
	$n = 21$; $p = 0.0986$; $X_3^2 = 1.51$; $P = 0.68$.		$n = 9$; $p = 0.2070$; $X_2^2 = 0.37$; $P = 0.83$.		$n = 10$; $p = 0.2142$; $X_3^2 = 0.91$; $P = 0.82$.	

	Text 10: Brief 9,9		Text 11: Brief 1,9		Text 12: Brief 2,6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	45	38.56	51	49.70	58	53.67
2	36	50.02	62	62.51	65	75.27
3	44	37.08	45	47.16	66	58.65
4	19	17.18	31	23.72	26	27.42
5	4	5.09	3	8.35	8	7.69
6	1	1.07	2	2.56	1	1.30
	$n = 8$; $p = 0.2704$; $X_3^2 = 6.73$; $P = 0.08$.		$n = 11$; $p = 0.2009$; $X_1^2 = 0.19$; $P = 0.66$.		$n = 7$; $p = 0.3186$; $X_3^2 = 2.82$; $P = 0.42$.	

	Text 13: Brief 9,30		Text 14: Brief 9,23		Text 15: Brief 8,16	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	35	32.99	52	47.01	44	38.64
2	43	41.31	58	69.01	44	56.88
3	24	31.04	62	54.03	53	44.65
4	22	15.54	21	23.79	20	19.71
5	1	5.45	7	6.16	4	5.12
6	3	1.67	-	-	-	-
	$n = 11$; $p = 0.2003$; $X_1^2 = 2.29$; $P = 0.13$.		$n = 6$; $p = 0.3700$; $X_2^2 = 3.91$; $P = 0.14$.		$n = 6$; $p = 0.3706$; $X_2^2 = 5.46$; $P = 0.07$.	

	Text 16: Brief 9,6		Text 17: Brief 8,22		Text 18: Brief 8,24		
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	NP_x*
1	32	31.35	35	31.47	111	100.63	109.64
2	46	46.56	41	47.22	108	122.02	110.47
3	40	38.41	39	37.79	86	92.06	89.01
4	16	19.01	19	17.01	64	48.37	53.78
5	4	5.64	3	4.08	18	18.77	21.66
6	2	0.93	1	0.43	2	7.15	4.39
7	2	0.10	-	-	-	-	-
	$n = 7$; $p = 0.3311$; $X_2^2 = 0.84$; $P = 0.66$.		$n = 6$; $p = 0.3751$; $X_2^2 = 1.54$; $P = 0.46$.		PN_x $n = 16$; $p = 0.1392$; $X_2^2 = 9.47$; $P = 0.0088$; $C = 0.0238$.	PN_x* $a = 0.2014$; $R = 6$; $X_3^2 = 4.01$; $P = 0.26$.	

	Text 19: Brief 6,16	
x	n_x	NP_x
1	128	125.92
2	219	224.79
3	223	214.02
4	110	114.61
5	30	32.73
6	6	3.93
	$n = 6$; $p = 0.4166$; $X_3^2 = 2.11$; $P = 0.55$.	

5. Die Untersuchung hat damit gezeigt, daß die Annahme, daß Wortlängen in Texten (stochastischen) Gesetzen gehorchen, auch im Falle der Plinius-Briefe bestätigt werden kann. Damit ist allerdings noch nichts für das Lateinische insgesamt ausgesagt: Ob die beiden gefundenen Modelle für andere Autoren, andere Zeitstufen und verschiedene Textsorten gelten, können nur weitere Untersuchungen zeigen. Dabei stößt man im Falle älterer Sprachen auf besondere Probleme der Textüberlieferung: Bevor man eine Textgruppe untersucht, muß man sich vergewissern, ob diese Texte zuverlässig tradiert wurden. Ist das nicht der Fall, wurden die ursprünglichen Texte also von späteren Abschreibern, Bearbeitern, Druckern etc. verändert, treten Probleme mit der Texthomogenität auf, die dazu führen können, daß völlig andere oder auch gar keine Modelle gefunden werden. Die Plinius-Briefe zeigen aber, daß alte Texte zumindest in diesem Fall recht gut tradiert wurden und offenbar keine Homogenitätsprobleme aufweisen.

Mit einem Homogenitätstest mit Hilfe der Informationsstatistik 2 I läßt sich leicht zeigen, daß sich für die Wortlängenhomogenität bei allen 19 Texten 2 I = 128.16 mit 102 Freiheitsgraden und $P = 0.04$ ergibt, was eine etwas kritische Homogenität andeutet. Läßt man aber die beiden kritischen Texte 8,14 und 8,24 aus, dann erhält man 2 I = 101.30 mit 90 Freiheitsgraden, was mit $P = 0.1952$ eine starke Homogenität signalisiert. Die abweichende Beschaffenheit der beiden Texte müßte philologisch oder mit anderen Methoden untersucht werden.

Literatur:

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio linguae II* (S. 19-30), Stuttgart: Steiner.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax*. 2. überarbeitete Auflage. Frankfurt: Athenäum.

C. Plini Caecili Secundi epistularum libri X recognovit brevique adnotatione critica instruxit R.A.B. Mynors. 2. überarbeitete Auflage. Oxford: Oxford UP 1982.

Rubenbauer, H., & Hofmann, J.B. (1977). *Lateinische Grammatik*. Neubearbeitet von R. Heine. Bamberg: Buchners Verlag u. a.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Glottometrika 16, 1997, 127-137

Wortlängenhäufigkeiten in italienischen Pressetexten

Cecilie Hollberg

1. Vorbemerkung

In dieser Arbeit geht es, wie schon in Gaeta (1995), darum, italienische Texte daraufhin zu untersuchen, welchen Verteilungsmodellen ihre Wortlängenhäufigkeiten folgen. Dazu wurden in diesem Fall ausschließlich Pressetexte bearbeitet. Die 10 Artikel dieser Untersuchung sind den drei folgenden italienischen Tageszeitungen entnommen:

1. „La Repubblica" ist die meistgelesene Tageszeitung in Italien; sie läßt sich in etwa mit „Die Welt" vergleichen.
2. „Il Giornale"; hierbei handelt es sich um eine 'großbürgerliche' Zeitung aus der politischen Mitte.
3. „Il Messaggero" ist eine römische Lokalzeitung mit relativ großer regionaler Bedeutung; sie behandelt natürlich vorwiegend römische Belange.

Bei den bearbeiteten Zeitungsartikeln wurden vorwiegend solche ausgewählt, die möglichst nur die italienischen Belange betreffen, um eine Überfremdung der Sprache zu umgehen. Als vergleichende Ausnahme dient ein Text der Rubrik „Ausland" (in diesem Falle Deutschland im Artikel 6).

Der Umfang der untersuchten Artikel beläuft sich auf eine Gesamtwortzahl von 625 bis 1280.

2. Definitionen

In dieser Untersuchung geht es darum, festzustellen, mit welcher Häufigkeit Wörter unterschiedlicher Länge in abgeschlossenen Texten vorkommen und ob es Verteilungen gibt, die die gefundenen Daten angemessen modellieren. Als Wort wird das orthographische „Wort" bestimmt (Bünting & Bergenholtz, 1989:36); die Länge der Wörter wird durch die Zahl ihrer Silben bestimmt. Als Kriterium für „Silbe" gilt das Vorhandensein von Vokalen oder Diphthongen.

Insofern folgt diese Untersuchung den in Best & Zhu (1994:20) genannten Prinzipien. Zur Anwendung auf italienische Texte sei noch auf folgende Besonderheiten hingewiesen: die Grapheme im Italienischen sind exakt die gleichen wie im Deutschen.

2.1 Diphthonge

Die Diphthonge im Italienischen sind eine silbische Einheit, die dadurch entsteht, daß ein <i> oder <u>, ohne betont zu sein, mit einem anderen Vokal in Verbindung steht; dieser andere Vokal kann betont oder unbetont sein[1].

Im Italienischen kann auch eine Folge von drei Vokalen zusammengehören, ohne zertrennt werden zu dürfen; diese nennen sich „trittongo", wie beispielsweise in den Wörtern „suoi" und „guai".

Ausnahmen stellen die Vokale dar, die einen Hiatus bilden.[2] Dieses gilt für drei Fälle. Zum einen bei der Folge von zwei Vokalen, von denen keiner ein <i> oder <u> ist. Die zweite Möglichkeit besteht, wenn die Betonung auf dem <i> respektive dem <u> liegt. Der letzte Fall hängt mit dem Präfix „ri-"[3] zusammen, da man hier eine Trennung der beiden Wortbestandteile deutlich heraushört.

Ebenso wird auch bei den Präfixen „bi-" und „tri-" verfahren, sofern sie sich von den Zahlen zwei und drei herleiten lassen, so z.B. bei „bi-ennio" (alle zwei Jahre) und „tri-angolo" (Drei-eck).

2.2 Die Silbe

Die Silbe in der italienischen Sprache ist ein Phonem oder eine Gruppe von Phonemen, welches sich in getrennter und autonomer Weise mit einer einzigen Lautemission artikulieren läßt, vgl. Dardano & Trifone (1989:332). Die Silbe wird immer mit einem Vokal gebildet. Auch einzelne Vokale können Silben darstellen; es gibt im Italienischen vier Wörter, die nur aus einem einzigen Vokal bestehen: „a"; „e"; „i"; „o" (übersetzt: zu, nach u.v.m.; und; Plural des Artikels „il", d.h.: die; oder).

Man unterscheidet zwischen „offenen" („aperte" oder „libere") Silben und geschlossenen („chiuse" oder „implicate"). Die offenen Silben enden mit einem Vokal, wie z.B. in dem Wort „te-le-fo-no", welches folglich aus vier offenen Silben besteht. Die geschlossenen Silben dagegen enden mit einem Konsonanten, wie die ersten drei Silben des Wortes „im-por-tan-za". Es gibt natürlich wie auch im Deutschen ein- und mehrsilbige Wörter.[4]

[1] Hierbei wird noch zwischen steigenden (ascendenti) und fallenden (discendenti) Diphtongen unterschieden, je nachdem ob die Betonung auf dem hinteren oder vorderen Teil des Diphtongs liegt.

[2] Der Hiatus in der italienischen Sprache muß nicht wie im Deutschen durch Zusammenfallen zweier Vokale am Wortanfang bzw. -ende stehen (dort würde er apostrophiert), sondern er kann mitten im Wort vorkommen und getrennt werden.

[3] „Ri-" bedeutet soviel wie „wieder, noch einmal".

2.3 Das Wort

Die Abkürzungen im Italienischen haben sehr einfache Ausspracheregeln. Alles, was „aussprechbar" ist, wie z.B. „USA", wird wie ein ganz normales Wort gelesen, sprich: /uza/, während Abkürzungen, die aus Konsonanten bestehen und daher nicht phonotaktische Wörter des Italienischen darstellen, buchstabiert gelesen werden, so z.B. „Dc", /ditʃi/. Die vollständige Lesart „Democrazia cristiana" oder Entsprechendes bei anderen Abkürzungen ist nicht berücksichtigt, da die normale Sprechweise auf die oben genannten Abkürzungen zurückgreift.

Namen, wie „Dalla Chiesa" sind als zwei getrennte Wörter gezählt, ebenso auch Doppelnamen, wenn sie auch durch einen Bindestrich verbunden sind, wie bspw. „Wieczorek-Zeul".

Auch durch Bindestrich zusammengefügte Wörter sind einzeln gezählt, wie „democristiano-socialdemocratico". Dazu zählen ebenfalls Beinamen wie beispielsweise: „Roberto il Guiscardo".

Feststehende Begriffe, die durch Bindestrich verbunden sind, sind dagegen als eine Einheit gezählt; so ist der Ausdruck „Centro-Nord" als ein 3-silbiges Wort gezählt.

2.4 Bearbeitung der Texte

Bei den vorliegenden Daten sind die Überschriften nicht mitberücksichtigt. Die Begründung hierfür liegt darin, daß eine Überschrift nicht unmittelbar mit dem Text zusammengehört; zumindest in Zeitungen stellen sie keinen Bestandteil des Textes dar, sondern wollen in schlagzeilenartiger Kurzzusammenfassung den Inhalt des Textes wiedergeben. Somit handelt es sich um eine Wiederholung dessen, was meist wörtlich im Text wiederkehrt. Diese Form der doppelten Aussage würde die Daten verfälschen. Satzzeichen sind nicht mitgezählt, sondern es wurde ausschließlich der laufende Text in seiner „hochitalienisch" gesprochenen Form ausgezählt.

[4] Man nennt diese „monosillabi" bzw. „polisillabi", die wiederum in „bi-, tri-, quadrisillabi..." unterteilt sind.

3. Modellierung

Die 10 ausgewerteten Pressetexte wurden mit dem Altmann-Fitter daraufhin untersucht, welche Modelle sich an sie anpassen lassen. Es zeigte sich, daß an die Texte die Palm-Poisson-Verteilung und die Cohen-Poisson-Verteilung, beide in 1-verschobener Form, angepaßt werden können, deren Formeln wie folgt lauten:

$$P_{x1} = \frac{R_{(x-1)}\, a^{x-1}}{F(R)}, \quad x = 1,2,\ldots,R+1$$

mit $n_{(x)} = n\,(n\text{-}1)\,\ldots\,(n\text{-}x\text{+}1)$ und $F(R) = \sum_{j=0}^{R} R_{(j)}\, a^{j}$;

$$P_{x2} = \begin{cases} e^{-a}(1+a\alpha), & x = 1 \\ ae^{-a}(1-\alpha), & x = 2 \\ \dfrac{e^{-a}a^{x-1}}{(x-1)!}, & x = 3,4,\ldots \end{cases}$$

Die beiden Modelle folgen aus dem allgemeinen Ansatz von Wimmer et al. (1994).

4. Die Daten der Pressetexte

Die Resultate der Anpassung sind in den unten aufgeführten Tabellen dargestellt. Hier bedeutet:

X^2	- der empirische Wert des Chiquadrat-Tests,
P	- die Irrtumswahrscheinlichkeit des Chiquadrats,
C	- der Diskrepanzkoeffizient: $C = X^2/N$,
a, R, α	- Parameter der Verteilungen,
x	- Silbenzahl pro Wort,
n_x	- beobachtete Häufigkeit der Länge x,
NP_{x1}	- theoretische Häufigkeit nach der Palm-Poisson-Verteilung,
NP_{x2}	- theoretische Häufigkeit nach der Cohen-Poisson-Verteilung.

Eine Anpassung betrachten wir als akzeptabel, wenn $P \geq 0{,}01$, oder, besonders bei großen Stichprobenumfängen, $C \leq 0{,}02$.

	Text 1		Text 2		Text 3	
x	n_x	NP_{x1}	nx	NP_{x1}	n_x	NP_{x1}
1	254	244.15	379	358.60	331	318.54
2	185	195.72	270	300.29	245	262.36
3	129	130.75	202	209.55	177	172.87
4	73	69.87	146	116.98	84	85.43
5	22	28.00	41	48.98	28	28.14
6	12	7.48	11	13.67	7	4.66
7	2	1.03	1	1.93	-	-
Σ	677	-	1050	-	872	-
	$a = 0.1336$; $R = 7$; $X_3^2 = 6.02$; $P = 0.11$; $C = 0.0088$.		$a = 0.1396$; $R = 7$; $X_4^2 = 13.94$; $P = 0.01$; $C = 0.0131$.		$a = 0.1647$; $R = 6$; $X_3^2 = 2.96$; $P = 0.40$; $C = 0.0034$.	

	Text 4		Text 5.		Text 6	
x	n_x	NP_{x1}	n_x	NP_{x2}	n_x	NP_{x1}
1	273	256.62	440	437.63	162	182.29
2	178	197.49	366	374.17	190	171.93
3	122	126.65	291	279.38	146	135.14
4	73	64.98	112	115.69	89	84.97
5	22	25.00	40	35.93	33	40.07
6	9	6.41	3	8.92	7	12.59
7	1	0.85	0	1.84	2	2.01
8	-	-	0	0.32	-	-
9	-	-	1	0.05	-	-
10	-	-	0	0.00	-	-
11	-	-	1	0.07	-	-
Σ	678		1254		629	
	$a = 0.1283$; $R = 7$; $X_3^2 = 5.55$; $P = 0.14$; $C = 0.0081$.		$a = 1.2423$; $\alpha = 0.1681$; $X_4^2 = 5.21$; $P = 0.27$; $C = 0.0041$.		$a = 0.1572$; $R = 7$; $X_4^2 = 8.96$; $P = 0.06$; $C = 0.0140$.	

	Text 7		Text 8		Text 9	
x	n_x	NP_{x2}	n_x	NP_{x1}	n_x	NP_{x1}
1	224	227.18	257	240.01	287	300.45
2	152	176.74	176	203.47	246	228.22
3	156	135.26	142	143.74	145	144.46
4	60	54.96	100	81.23	81	73.15
5	20	16.75	32	34.43	15	27.78
6	0	4.08	6	9.73	6	7.03
7	1	0.83	1	1.39	2	0.91
8	0	0.14	-	-	-	-
9	0	0.02	-	-	-	-
10	0	0.00	-	-	-	-
11	3	0.04	-	-	-	-
Σ	616	-	714	-	782	-
	$a = 1.2191$; $\alpha = 0.2035$; $X_3^2 = 8.01$; $P = 0.09$; $C = 0.0128$.		$a = 0.1413$; $R = 7$; $X_4^2 = 10.97$; $P = 0.03$; $C = 0.0151$.		$a = 0.1266$; $R = 7$; $X_3^2 = 8.71$; $P = 0.03$; $C = 0.0110$.	

	Text 10	
x	n_x	NP_{x2}
1	266	275.99
2	240	236.99
3	174	158.66
4	64	61.08
5	10	17.63
6	1	5.04
7	-	-
Σ	755	-
	$a = 1.1550$; $\alpha = 0.1374$; $X_3^2 = 8.51$; $P = 0.04$; $C = 0.0111$.	

Text 1 Amministrative, andremo due volte alle urne per eleggere il sindaco (Verwaltung, wir werden zweimal zu den Urnen gehen, um den Bürgermeister zu wählen) . In: „Il Messagero", 24.5.1993, S.2, (Politik).

Text 2 Andreotti, oggi si apre l'armadio dei misteri (Andreotti, heute öffnet sich der Schrank der Mysterien). In: „La Repubblica", 14.4.1993, S.3, (Politik).

Text 3 Nasce la loggia dei ribelli - il Gran maestro spacca i massoni (Die Loge der Rebellen ist entstanden- der Großmeister spaltet die Freimaurer). In: „La Repubblica", 18/19.4.1993, S. 17, (Politik).
Anmerkungen[5]: „Centro-Nord" ist als ein Wort gezählt, da es ein feststehender Ausdruck ist. „Di Bernardo" ist zwar ein zusammengehöriger Name, der aber eindeutig aus zwei verschiedenen Wörtern besteht, daher auch als zwei Wörter gewertet. Ein lateinischer Satz kommt im Text vor: „Jure veritati juncti"; er ist mitgezählt.

Text 4 Se ci fosse uno stato (Wenn es einen Staat gäbe). In: Ebda., S. 9.

Text 5 Noi clochard dell'arco (Wir Clochards vom <Janus->Bogen). In: „Il Giornale", 29.7.1993, S. 8, (Politik).
Anmerkungen: Das 9- bzw. 11-silbige Wort ist in beiden Fällen eine Jahreszahl. Nicht-italienische Namen und Begriffe: „Gauleiter; Tischbein; Ruiz; Goethe; Stendhal".

Text 6 Grande coalizione ? No ma tutta Bonn ne parla (Große Koalition ? - Nein, aber ganz Bonn redet davon). In: „La Repubblica", 24.8.1993, S. 13. (Außenpolitik).
Anmerkungen: In diesem Artikel sind ausgesprochen viele deutsche Namen, sie sind mit italienischer Betonung gezählt, was sich nur bei „Mueller" bemerkbar macht, da er im Italienischen „Mu-el-ler" (3-silbig) gesprochen wird. Weitere deutsche Wörter: „Sueddeutsche Zeitung; Laender; Bundesrat". Bei „Sueddeutsche" sprechen die Italiener das <ü> wie <u>, es kommt daher zu keiner Abweichung in der Silbenzählung. „Democristiano-socialdemocratico" ist als zwei Wörter gezählt, da es kein feststehender Ausdruck ist.

Text 7 Anche gli scugnizzi avranno un santo (Auch die Lausebengels werden einen Heiligen haben). In: „La Repubblica", 18/19.4.1993, S. 20, (Berichterstattung).
Anmerkungen: „Scugnizzi" ist ein Wort aus dem neapolitanischen Dialekt, ist, und Der Name „Casoria" ist hier mit 3 Silben gezählt, sofern die

[5] Unter: „Anmerkungen" ist alles aufgeführt, was von der normalen Zählung abweicht, bzw. was an Auffälligkeiten oder Besonderheiten zu dem jeweiligen Artikel anzumerken ist.

Betonung auf dem <o> liegt; legte man sie auf das <i>, würden es 4 Silben werden. „71 anni" würde korrekterweise „settantuno anni" gesprochen, in der Praxis spricht man es aber „settantun' anni", womit eine Silbe verloren geht, so ist es auch hier gezählt. Die 11-silbigen Wörter sind wieder Jahreszahlen. Das 7-silbige lautet „indaffaratissimo", es ist ein Superlativ mit dem Morphem „-issimo"; häufig sind diese Endungsmorpheme dafür verantwortlich, daß die Wörter so lang sind. Der neapolitanische Satz in Artikel 7 führt zu einer exakten Verdoppelung der Wörter, da zwar die Aussprache eine andere ist, nicht aber die Silbenzahl.

Text 8 E' una favola o la realtà ? (Ist es ein Märchen oder die Realität ?). In: Ebda., S. 30, (Kultur: Musik).

Text 9 I ragazzi contro il Muro (Die Jungs gegen die Mauer) In: „La Repubblica", 14.4.1993, S. 41, (Sport).
Anmerkungen: Viele Namen verschiedenster Nationalitäten. Mercato Uno" ist als 5-silbiges Wort gezählt.

Text 10 Treviso bella rabbia (Treviso, schöner Ärger). In: Ebda, S. 40, (Sport).
Anmerkungen: Viele Namen, und vor allem viele Zahlen. Die Lesart von <2"> ist „due secondi"; <2'> hingegen „due minuti". Das „1 + 1" wird „uno più uno" gelesen. Im letzten Absatz werden die verschiedenen Spieler mit ihren Punkten angegeben, hier werden die Kommata zwischen den Zahlen mitgelesen: „7,5" lautet folglich: „sette virgola cinque".

5. Namen

Verständlicherweise fallen in Zeitungen viele Namen von Städten und Personen, die nicht immer der Nationalität der Zeitungen entsprechen, wobei sich die Frage aufdrängt, ob durch die fremden Namen nicht die Silbenzahlen verfälscht werden. In diesem Abschnitt ist ein Großteil der Namen, die in den Artikeln vorkamen, auf ihre Silbenzahl hin untersucht worden. Der Beweggrund dafür ist es festzustellen, ob und inwieweit die Silbenzahlen der italienischen Namen von denen aus anderen Sprachen abweichen. In der folgenden Tabelle sind die Namen nach Silbenzahlen sowie italienischer bzw. internationaler Zugehörigkeit aufgeführt.

Als Ergebnis ist also festzuhalten, daß von 6 1-silbigen Namen keiner italienisch ist. Von den insgesamt 31 2-silbigen Namen sind immerhin schon 12 italienisch. Während unter den 3-silbigen Wörtern erstaunlicherweise von 34 Namen 25 aus Italien kommen und nur 9 aus anderen Ländern.

1-silbig	international	Rooks, Kohl, Hans, Zeul, Klaus, Bob **Total: 6**
2-silbig	italienisch	(Lo) Forte, Ciampi, Indro, Gino, Giofà, Fuochi, Raffi, Corti, Claudio, Bugno, Giulio, Gianni **Total: 12**
	international	Dylan, Kukok, Teagle, Korfas, Skansi, Barlow, Furlan, (de) Vlaeminck, Richard, Ulrich, Gerhard, Schröder, Albrecht, Theo, Waigel, Helmut, Scharping, Rudolf, Kinkel **Total: 19**
3-silbig	italienisch	Corona, Armando, Cordona, (di) Bernardo, Piccoli, Bartoli, Chioccioli, Cassari, Chiappucci, Vallone, Giuliano, (de) Simone, Ambrogio, Sparagna, Castaldo, Lucilla, Rusconi, Vianini, Ragazzi, (di) Giacomo, Francesco, Achille, Torelli, Mancuso, Navarro **Total: 25**
	international	Fondriest, Argentin, Criquielion, Balladur, Wieczorek, Levingston, Prelevic, Fassoulas, Bourdouris **Total: 9**
4-silbig	italienisch	Casagrande, Andreotti, Iacobacci, Ferdinando, Galeazzi, Montanelli, Ludovico **Total: 7**
	international	Indurain, Heidemarie **Total: 2**

Die Mehrzahl der 4-silbigen Namen liegt auch bei den Italienern, mit 7 zu nur 2 anderen Namen. Die Anteile italienischer Namen bei einzelnen Wortlängen sind wie folgt:

Länge	1	2	3	4
%	0	39	74	78

worin man auch ohne Test einen klaren Trend erkennt.

Dieses Ergebnis ist erstaunlich, da die italienische Sprache im Verhältnis z.B. zur deutschen kürzere Wörter hat; warum es bei den Namen umgekehrt ist, bleibt vorerst ungeklärt.

6. Schlußbemerkungen

Die Untersuchung hat ergeben, daß alle bearbeiteten Pressetexte mit Wahrscheinlichkeitsverteilungen modelliert werden können. Wie schon bei Gaeta (1995) zeigte sich jedoch, daß verschiedene Modelle berücksichtigt werden müssen. Schon unauffällige Pressetexte erwiesen sich damit im Italienischen als weniger homogen als z.B. die deutschen (Best, 1997) oder die französischen (Dieckmann & Judt, 1996). Wie schon in Gaetas Untersuchung (Gaeta, 1995) bewährt sich aber auch hier die Palm-Poisson-Verteilung als ein grundlegendes Modell der Wortlängenverteilungen in italienischen Texten. Im Vergleich mit Gaetas Resultaten zeigt sich eben, daß im Italienischen die Wortlänge keine einheitliche Tendenz hat, sondern sich zeit-, genre- und autorenabhängig gestaltet. Man kann erwarten, daß weitere Untersuchungen eine noch stärkere Modellversifikation mit sich bringen werden.

Erstaunlich ist das Ergebnis des Vergleichs italienischer mit fremdsprachigen Namen: Während die Appellativa im Italienischen keine auffällige Tendenz zu höherer Wortlänge zeigen, scheinen italienische Eigennamen im Verhältnis zu denen anderer Sprachen durchschnittlich länger zu sein. Natürlich müßte dieser Befund an wesentlich umfangreicherem Material abgesichert werden, als das hier geschehen konnte.

Literatur

Akademie-Grammatik: Grundzüge einer deutschen Grammatik. Von einem Autorenkollektiv unter der Leitung von K.E. Heidolph, W. Fläming & W. Motsch. Berlin: Akademie Verlag, 1981.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In diesem Band,

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Bünting, K.-D., & Bergenholtz, H. (^{2}1989). *Einführung in die Syntax*. Frankfurt: Athenäum.

Dardano, M., & Trifone, P. (1989). *Studiamo la lingua - elementi di grammatica italiana*. Firenze: Zanichelli.

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In P. Schmidt (Hg.) *Glottometrika 15* (S. 158-165), Trier: WVT.

Gaeta, L. (1995). Wortlängenverteilung in italienischen Texten. *Zeitschrift für empirische Textforschung*, 1, 44-48.

Il Giornale, 29.7.1993.

Il Messagero, 24.5.1993.

La Repubblica, 14.4.1993; 18/19.4.1993; 24.8.1993.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Zingarelli, N. (1991). *Il nuovo Zingarelli - vocabolario della lingua italiana*. Bologna: Zanichelli.

SOFTWARE

Altmann-FITTER. Lüdenscheid: RAM-Verlag, 1994.

Glottometrika 16, 1997, 138-144

Wortlängen in Briefen des spanischen Dichters Federico García Lorca

Martina Hein

Die vorliegende Arbeit entstand im Zusammenhang mit dem Göttinger und Bochumer Projekt zur Wortlängenforschung (vgl. entsprechende Beiträge in Glottometrika 15 und 16), das bisher knapp 30 Sprachen untersucht hat. Dabei stellte sich heraus, daß für alle Sprachen Modelle dafür gefunden werden konnten, wie sich die Häufigkeit, mit der die unterschiedlichen Wortlängen in den einzelnen Texten auftreten, darstellt.

Im folgenden geht es darum, entsprechende Daten für das Spanische zu präsentieren. Untersuchungsgegenstand sind 20 Briefe des spanischen Dichters Federico García Lorca, die er zwischen 1918 und 1921 an Freunde, Dichter und ihm nahestehende Frauen geschrieben hat. Briefe bieten sich als Textsorte für eine Untersuchung besonders an, da sie in der Regel von einem Bearbeiter stammen, der ohne zeitliche Unterbrechung einen homogenen Text verfaßt (vgl. Bartels & Sehlow, 1997).

Um eine Analyse von Texten nach der Häufigkeit von Wörtern unterschiedlicher Länge, gemessen in der Zahl der Silben pro Wort, durchführen zu können, müssen die Einheiten „Wort" und „Silbe" bestimmt werden. Wie schon in Best und Zhu (1994) wird das orthographische Wort als Untersuchungseinheit gewählt. Die Silbe wurde nach dem Vorkommen von Vokalen und Diphthongen bestimmt, d.h. ein Wort enthält soviele Silben wie Vokale oder Diphthonge.

Hierbei ist für das Spanische folgendes zu beachten: Eine allgemeine Regel besagt, daß Diphthonge durch einen Akzent gebrochen werden, d.h. zwei getrennte Silben bilden. Ein Beispiel wäre „día", das durch den Akzent zweisilbig wird. Allerdings werden die Diphtonge nur dann als zwei einzelne Laute gesprochen, wenn ein sogenannter „schwacher" Vokal („u" oder „i") den Akzent trägt. Folgen zwei starke Vokale aufeinander, wird ein sogenannter Hiat gebildet (ebenfalls eine Trennung in zwei Silben). Ein Beispiel dafür wäre das dreisilbige Verb „contraer".

Abweichend von Sünje von Ahn und Karl-Uwe Potthast (1993), die 17 Pressetexte aus Ecos de España, El País und Panorama (darunter ein Leserbrief)

sowie vier Kurzgeschichten verschiedener Autoren ausgewertet haben, habe ich das Verb „fue" einsilbig gezählt, da der Diphthong, der aus einem „starken" Vokal („e") und einem schwachen („u") besteht, keinen Akzent trägt und folglich keine silbische Trennung vorliegt. Die Ergebnisse der Untersuchung haben gezeigt, daß trotz der unterschiedlichen Art der Datenerhebung das gleiche Modell, die 1-verschobene gemischte Poisson-Verteilung, an die Daten angepaßt werden konnte, die wie folgt lautet:

$$P_x = \frac{\alpha\, a^{x-1} e^{-a}}{(x-1)!} + \frac{(1-\alpha)\, b^{x-1} e^{-b}}{(x-1)!}, \quad x = 1, 2, 3, \ldots$$

Die Anpassung der 1-verschobenen gemischten Poisson-Verteilung an die Daten der 20 Briefe Frederico García Lorcas hat die folgenden Ergebnisse erbracht:

In den Tabellen bedeutet:

x	- die Wortlängenklasse;
n_x	- die beobachtete Häufigkeit, mit der die Wortlängenklasse im jeweiligen Text vorkommt;
NP_x	- die aufgrund der 1- verschobenen gemischten Poisson-Verteilung berechneten Werte der Häufigkeit;
X^2	- das Chiquadrat;
FG	- die Freiheitsgrade;
P	- die Wahrscheinlichkeit, das festgestellte Chiquadrat zu überschreiten;
a, b, c	- die Parameter des Modells.

Die Berechnungen wurden mit dem Altmann-Fitter (1994) durchgeführt. Die Anpassung des Modells an die Daten ist zufriedenstellend, wenn P größer oder gleich 0.05 ist. Von den Briefen wurde weder der Briefkopf (Anrede und Datum) noch die Unterschrift ausgewertet.

	Brief 1		Brief 2		Brief 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	120	119.77	208	207.04	77	75.23
2	83	82.43	142	140.83	59	57.86
3	44	45.74	77	84.19	27	27.40
4	22	20.76	46	37.64	6	10.95
5	7	7.46	12	12.82	8	5.58
6	3	2.84	1	3.50	-	-
7	-	-	0	0.80	-	-
8	-	-	1	0.22	-	-
	$a = 1.4602$; $b = 0.3124$; $\alpha = 0.6054$; $X_2^2 = 0.18$; $P = 0.91$.		$a = 1.3650$; $b = 0.1681$; $\alpha = 0.7122$; $X_2^2 = 3.91$; $P = 0.14$.		$a = 1.6709$; $b = 0.6173$; $\alpha = 0.3255$; $X_1^2 = 3.37$; $P = 0.07$.	

1. Brief vom 1. Februar 1918.
2. Brief vom 19.September 1920.
3. Brief vom November 1919.

	Brief 4		Brief 5		Brief 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	170	167.88	124	124.54	113	112.47
2	95	95.87	99	96.90	78	78.38
3	53	58.32	44	47.49	49	47.93
4	32	24.12	23	20.33	19	20.49
5	5	7.48	8	7.93	6	6.59
6	1	2.34	2	2.73	3	2.16
7	-	-	0	0.81	-	-
8	-	-	1	0.30	-	-
	$a = 1.2412$; b = 0.0208; $\alpha = 0.7356$; $X_2^2 = 4.67$; $P = 0.97$.		$a = 1.8261$; $b = 0.6281$; $\alpha = 0.3217$; $X_3^2 = 0.86$; $P = 0.84$.		$a = 1.2865$; $b = 0.0749$; $\alpha = 0.7799$; $X_2^2 = 0.54$; $P = 0.76$.	

4. Brief vom Juni 1923.
5. Brief vom November 1920.
6. Brief vom Frühling 1921.

	Brief 7		Brief 8		Brief 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	157	156.39	91	89.68	55	53.96
2	86	86.96	61	61.25	42	39.82
3	51	52.20	34	38.48	19	21.67
4	24	21.11	22	16.53	8	10.71
5	6	6.41	5	7.09	9	6.84
6	1	1.96	-	-	-	-
	$a = 1.2135$; $b = 0.0109$; $\alpha = 0.7340$; $X_2^2 = 0.91$; $P = 0.63$.		$a = 1.2899$; $b = 0.0376$; $\alpha = 0.7881$; $X_1^2 = 2.96$; $P = 0.09$.		a = 1.8141; $b = 0.4976$; $\alpha = 0.4545$; $X_1^2 = 1.84$; $P = 0.17$.	

7. Brief vom August 1921.
8. Brief vom August 1921.
9. Brief von 1922.

	Brief 10		Brief 11		Brief 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	229	228.70	58	57.15	129	128.62
2	156	156.19	32	31.71	82	81.38
3	102	103.25	17	20.39	46	47.66
4	49	48.01	13	9.81	24	24.20
5	18	16.79	4	4.94	12	9.80
6	4	4.70	-	-	2	3.21
7	1	1.38	-	-	1	1.16
	$a = 1.3992$; $b = 0.0735$; $\alpha = 0.7621$; $X_3^2 = 0.32$; $P = 0.96$.		$a = 1.4578$; $b = 0.1056$; $\alpha = 0.6580$; $X_1^2 = 1.79$; $P = 0.18$.		$a = 1.6418$; $b = 0.2950$; $\alpha = 0.5628$; $X_3^2 = 1.03$; $P = 0.79$.	

10. Brief vom 1. Juli 1922.
11. Brief vom Juli 1922.
12. Brief von Dezember 1922

	Brief 13		Brief 14		Brief 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	144	142.13	92	91.58	83	82.94
2	74	75.32	64	65.85	62	60.09
3	50	45.70	35	32.55	27	30.78
4	16	23.47	12	10.74	15	13.02
5	11	9.35	0	2.66	5	4.50
6	5	4.04	1	0.66	1	1.68
	$a = 1.6020$; $b = 0.1921$; $\alpha = 0.5635$; $X_2^2 = 3.36$; $P = 0.19$.		$a = 0.9897$; $b = 0.0029$; $\alpha = 0.8765$; $X_1^2 = 1.98$; $P = 0.16$.		$a = 1.4404$; $b = 0.4211$; $\alpha = 0.5402$; $X_2^2 = 1.15$; $P = 0.56$.	

13. Brief vom Juli 1924.
14. Brief vom September 1924.
15. Brief vom September 1925.

	Brief 16		Brief 17		Brief 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	79	78.41	64	63.53	66	64.56
2	55	57.46	60	57.98	43	44.05
3	41	35.10	27	32.11	27	30.15
4	11	14.31	17	12.99	19	13.82
5	3	4.38	3	4.08	4	6.45
6	2	1.35	1	1.32	-	-
	$a = 1.2230$; $b = 0.0018$; $\alpha = 0.8349$; $X_2^2 = 2.63$; $P = 0.27$.		$a = 1.2804$; $b = 0.4287$; $\alpha = 0.7552$; $X_2^2 = 2.49$; $P = 0.29$.		$a = 1.3747$; $b = 0.0056$; $\alpha = 0.7935$; $X_1^2 = 3.25$; $P = 0.07$.	

16. Brief vom Februar- März 1926.
17. Brief vom Oktober 1926.
18. Brief vom Februar 1927.

	Brief 19		Brief 20	
x	n_x	NP_x	n_x	NP_x
1	71	70.94	274	272.26
2	60	59.81	149	149.37
3	36	36.76	90	92.57
4	17	16.65	41	41.23
5	6	5.77	18	13.83
6	2	2.09	2	4.76
	$a = 1.3925$; $b = 0.2507$; $\alpha = 0.7716$; $X_2^2 = 0.04$; $P = 0.98$.		$a = 1.3423$; $b = 0.0711$; $\alpha = 0.6820$; $X_2^2 = 2.92$; $P = 0.23$.	

19. Brief vom Juli 1927.
20. Brief vom 5. April 1930.

Die Untersuchung der 20 Briefe Federico García Lorcas hat damit gezeigt, daß an alle Texte problemlos dasselbe Modell, die 1-verschobene gemischte Poisson-Verteilung, angepaßt werden kann.

Insgesamt läßt sich damit für das Spanische folgendes feststellen: Die 1-verschobene gemischte Poisson-Verteilung stellt offenbar ein gutes Modell für das gegenwärtige Schriftspanisch dar, da sich nur ein Text von insgesamt 41 nicht mit diesem Modell darstellen läßt. Eine mögliche Erklärung für diese Unregelmäßigkeit wäre eine Störung der Texthomogenität durch nachträgliche Bearbeiter (denkbar wären z.B. Veränderungen des Textes durch die Redaktion).

Auch wenn das untersuchte Textkorpus aus bisher nur 41 Texten besteht und das Ergebnis noch nicht verallgemeinert werden darf, so ist doch bemerkenswert, daß die verschiedenartigsten Textsorten (private sowie öffentliche) demselben Modell folgen. In einer weiteren Untersuchung werden weitere Daten für das Spanische erarbeitet, durch deren Auswertung die bisherigen Ergebnisse ein weiteres Mal überprüft werden.

Abschließend sei noch auf Forschungsperspektiven verwiesen, die sich mit dieser und ähnlichen Untersuchung eröffnen. Ein erster Aspekt ist der typologische. Fucks (1968:91) hat neben anderen Sprachen auch das Französische und Italienische mit ihren Wortlängen, gemessen an der Zahl der Silben pro Wort, erfaßt. Wenn für das Spanische ein repräsentatives Textkorpus erarbeitet worden ist, kann man auch diese Sprache in den Vergleich miteinbeziehen.

Ein weiterer Aspekt wurde von Wimmer, Best und Altmann (erscheint) bearbeitet. Diese Autoren versuchen, anhand der bekannten Daten aus dem Lateini-

schen, Französischem, Italienischen und Spanischen die Romania hinsichtlich ihrer Wortlängenentwicklung und synchronen Ausprägung zu charakterisieren; auch hier bedarf es noch weiterer Untersuchungen, um zu einem relativ vollständigen Überblick zu gelangen.

Literatur:

Ahn, S. von, & Potthast, U. (1993). Hausarbeit zur Silbenzählung in spanischsprachigen Texten.

Altmann-Fitter (1994). Lüdenscheid, RAM-Verlag.

Best, K.- H., & Zhu, J. (1994). Zur Häufung von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio linguae II* (S. 19-30), Stuttgart: Steiner.

Bartels, O., & Sehlow, M. (1997). Zur Häufigkeit von Wortlängen in deutschen Briefen im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts (Bismark, Brecht, Kafka, Thomas Mann, Tucholsky). In diesem Band

Colección obras eternas. Madrid:. Aguilar S.A. de Ediciones, 1986.

Poesía: Revista ilustrada de información poetica, N° 23-24. Federico García Lorca escribe a su familia desde Nueva York y la Habana (1929-1930).

Wimmer, G., Best, K.-H., & Altmann, G. (1996). Wortlänge in romanischen Sprachen. In *Festschrift für U. Figge* (erscheint).

Glottometrika 16, 1997, 145-151

Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten

Sabine Feldt, Marianne Janssen, Silke Kuleisa

0. Die vorliegende Untersuchung entstand im Rahmen eines Forschungsprojekts, das sich mit der Häufigkeit von Wortlängen in Texten befaßt. Es geht dabei darum, herauszufinden, ob auch für das Französische nachgewiesen werden kann, daß die Häufigkeit, mit der Wörter verschiedener Länge in Texten unterschiedlicher Art verwendet werden, Gesetzmäßigkeiten folgt, wie die Theorie (vgl. dazu z.B. Grotjahn, 1982; Wimmer et al., 1994) erwarten läßt, und wenn das der Fall sein sollte, welche Gesetzmäßigkeiten es sind, die die Verhältnisse im Französischen regeln. Für die Modellierung von Wortlängenhäufigkeiten steht eine Vielzahl von Funktionen zur Verfügung; die bisherigen Untersuchungen (z.B. Best & Zhu, 1994; Best, 1996, 1997; Nemcová & Altmann, 1994) haben gezeigt, daß in verschiedenen Sprachen und innerhalb der Sprachen in verschiedenen Funktionalstilen mit unterschiedlichen Modellen gerechnet werden muß. Dabei ist wichtig, daß gezeigt werden kann, daß eine Reihe von Funktionen lediglich Varianten einer gemeinsamen zugrundeliegenden Gesetzmäßigkeit darstellen (Wimmer et al., 1994; Wimmer & Altmann, 1996), die als Anpassung an verschiedene Randbedingungen interpretiert werden können.

1. Für diese Untersuchung wurden insgesamt sechs Pressetexte und 16 Briefe untersucht. Bei der Textbearbeitung wurden die gleichen Prinzipien berücksichtigt, die auch für andere derartige Untersuchungen gelten (vgl. Best & Zhu, 1994; Best, 1996), allerdings mit einigen Veränderungen, um den Besonderheiten des Französischen, vornehmlich der französischen Orthographie, gerecht werden zu können. Als Untersuchungseinheit wird das orthographische Wort aufgefaßt, ohne Rücksicht auf die „chaîne parlée"; die Zahl der Silben im Wort bestimmt sich nach der Zahl der Vokale, die im entsprechenden Wort enthalten sind. Vokale, die zwar geschrieben, aber nicht als Vokale gesprochen werden (z.B. <e> in „belle", <i> in „mieux"), gelten nicht als Silbenträger. In Zweifelsfällen wurde auf Handbücher (Petit Robert u.a.) zurückgegriffen.

Nullsilbige Wörter sind zunächst gesondert erhoben worden, wenn vokallose, apostrophierte Wörter wie „l'" (Artikel) vorkamen. Sie werden in den Tabellen aber nicht mehr eigens als Wortlängenklasse aufgeführt, sondern als phonetischer Bestandteil ihrer Nachbarwörter betrachtet. Erste Auswertungen haben nämlich gezeigt, daß Berechnungen unter Berücksichtigung der nullsilbigen Wörter zu dem gleichen Verteilungsmodell geführt haben, wie wenn man diese Wortlängenklasse nicht berücksichtigt.

Ein besonderes Problem stellen im Französischen die Wörter dar, die mit Bindestrich verbunden sind. Anders als z.B. im Deutschen kann man sie nicht immer als jeweils ein Wort auffassen: frei umstellbare, nicht lexikalisierte Formen wie „voulez-vous" etc. werden daher als Folgen von zwei Wörtern aufgefaßt; lexikalisierte Formen wie „tête-à-tête", „peut-être", „là-bas" dagegen sind Komposita und werden daher als jeweils nur ein Wort behandelt. Letzteres gilt auch für Eigennamen wie „Juan-les-Pins".

Abkürzungen wie „A.D.", „vs vs" (für „vous vous") werden ebenso wie Zahlwörter in ihrer gesprochenen Form gewertet. Bei der Textauswertung wurde immer der laufende Text berücksichtigt; bei den Zeitungstexten zusätzlich die Überschriften und bei den Briefen die Anrede und die Grußformel. Andere evtl. vorhandene Textteile wurden nicht ausgewertet.

2. Alle bisher untersuchten französischen Texte (vgl. auch Dieckmann & Judt, 1996; Wimmer & Altmann, 1996:129) folgen der Hirata-Poisson-Verteilung:

$$P_x = \sum_{i=0}^{[x/2]} \binom{x-1}{i} \frac{e^{-a} a^{x-1}}{(x-1)!} \alpha^i (1-\alpha)^{x-2i}, \quad x = 0,1,2,\ldots$$

Dabei sind α und a Parameter der Funktion. Außerdem werden in den Tabellen angegeben: x (Wortlängenklasse), n_x (Zahl der Wörter im Text mit entsprechender Silbenzahl), NP_x (nach der Hirata-Poisson-Verteilung berechnete Werte), X^2 (Chiquadrat), P (Überschreitungswahrscheinlichkeit des entsprechenden Chiquadrats). Die Anpassung wird als zufriedenstellend gewertet, wenn $P \geq 0{,}05$.

Die Ergebnisse stellen sich wie folgt dar:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	690	690.95	418	415.48	617	616.27
2	244	244.46	175	175.43	304	307.07
3	81	75.97	61	66.29	117	114.60
4	13	16.68	23	17.56	35	31.69
5	4	3.94	3	5.24	5	7.49
6	-	-	-	-	1	1.88
	$a = 0.4012$; $\alpha = 0.1181$; $X_2^2 = 1.15$; $P = 0.56$.		$a = 0.4927$; $\alpha = 0.1429$; $X_2^2 = 3.07$; $P = 0.22$.		$a = 0.5601$; $\alpha = 0.1104$; $X_3^2 = 1.66$; $P = 0.65$.	

Text 1: Frédérique Hébrard: La belle femme hors du temps (aus: „Madame Figaro - Jour de France", No. 9331 vom 31.7. - 6.8.1993, éd. internationale, S.5). Kurzgeschichte.

Text 2: Valery Bailly: Claudia Schiffer: La plus belle fille au monde (Textquelle wie Text 1, S. 24). Reportage.

Text 3: Stéphane Bern: Les mariés de l'été (Textquelle wie Text 1, S. 27ff.). Reportage.

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	325	323.07	273	270.69	210	208.89
2	144	144.82	134	135.53	101	101.97
3	76	79.45	61	65.03	56	58.10
4	29	25.91	27	21.23	23	20.26
5	10	8.68	6	8.52	9	9.78
6	1	3.07	-	-	-	-
	$a = 0.5937$; $\alpha = 0.2450$; $X_3^2 = 2.11$; $P = 0.55$.		$a = 0.6156$; $\alpha = 0.1867$; $X_2^2 = 2.58$; $P = 0.28$.		$a = 0.6471$; $\alpha = 0.2457$; $X_2^2 = 0.52$; $P = 0.77$.	

Text 4: Jean-Paul Picaper: Le nouvel homme fort des sociaux-démocrates (aus: Le Figaro 23.06.93, S.3, Rubrik/Sparte: La vie internationale). Vorwiegend informationsbetonter Text. Funktionalstil: Presse und Publizistik.

Text 5: Gérard Nirascou: École libre: l'obsolète loi Falloux (aus: Le Figaro 23.06.93, S.2, Rubrik/Sparte: Opinions). Vorwiegend meinungsbetonter Text. Funktionalstil: Presse und Publizistik.

Text 6: Jacques Malherbe: Les choix de l'après-bac (aus: Le Figaro 23.06.93, S.15, Rubrik/Sparte: Que faire après le bac?). Vorwiegend informationsbetonter Text. Funktionalstil: Presse und Publizistik.

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	620	617.82	267	264.19	303	301.64
2	197	196.84	80	80.66	111	107.09
3	56	61.41	23	27.68	18	25.80
4	16	12.90	9	5.94	8	5.47
5	0	2.52	1	1.53	-	-
6	3	0.51	-	-	-	-
	$a = 0.3673$; $\alpha = 0.1315$; $X_2^2 = 1.23$; $P = 0.54$.		$a = 0.3635$; $\alpha = 0.1600$; $X_2^2 = 2.57$; $P = 0.28$.		$a = 0.3775$; $\alpha = 0.0596$; $X_1^2 = 3.68$; $P = 0.05$.	

Text 7: Jean-Paul Sartre: Brief an Simone de Beauvoir, 11. Mai 1937 (aus: Beauvoir, Simone de (Hg.), *Lettres au Castor et à quelques autres [1926-39, Jean-Paul Sartre]*. Bd. 1. Paris, Gallimard, 1983, S. 147-149).

Text 8: Jean-Paul Sartre: Brief an Simone de Beauvoir, ohne Datum (vermutlich Juli/August 1939) (Textquelle wie Text 7, S. 243f.).

Text 9: Jean-Paul Sartre: Brief an Simone de Beauvoir, 27. September 1939 (Textquelle wie Text 7, S. 319f.).

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	529	530.27	204	203.80	163	162.86
2	180	179.98	94	95.23	65	64.37
3	83	79.77	36	34.29	22	22.63
4	19	20.16	10	9.09	5	5.59
5	6	6.82	1	2.59	1	1.24
6	-	-	-	-	1	0.31
	$a = 0.4322$; $\alpha = 0.2148$; $X_2^2 = 0.30$; P = 0.86.		$a = 0.5264$; $\alpha = 0.1123$; $X_2^2 = 1.15$; $P = 0.56$.		$a = 0.4562$; $\alpha = 0.1334$; $X_2^2 = 0.24$; $P = 0.89$.	

Text 10: Jean Grenier: Brief an Albert Camus, Ostern 1943 (aus: Grenier, Jean, *Correspondance, 1932-60*. Paris, Gallimard, 1981).

Text 11: Jean Grenier: Brief an Albert Camus, 7.1.1947 (Textquelle wie Text 10).

Text 12: Jean Grenier: Brief an Albert Camus, 20.11.1949 (Textquelle wie Text 10).

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	914	914.21	488	488.91	355	355.70
2	294	295.67	168	168.27	118	185.03
3	106	102.37	61	57.33	49	52.60
4	24	22.80	11	13.08	12	10.67
5	3	5.95	1	2.79	2	2.00
6	-	-	2	0.62	-	-
	$a = 0.3831$; $\alpha = 0.1558$; $X_2^2 = 1.65$; $P = 0.44$.		$a = 0.4022$; $\alpha = 0.1443$; $X_2^2 = 0.61$; $P = 0.74$.		$a = 0.5328$; $\alpha = 0.0236$; $X_2^2 = 0.46$; $P = 0.79$.	

Text 13: Albert Camus: Brief an Jean Grenier, 18.6.1938 (Textquelle wie Text 10).

Text 14: Albert Camus: Brief an Jean Grenier, Frühjahr 1940 (Textquelle wie Text 10).

Text 15: Albert Camus: Brief an Jean Grenier, 9.3.1943 (Textquelle wie Text 10).

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	347	347.05	341	339.76	458	452.45
2	136	137.71	130	130.64	173	173..00
3	51	43.67	44	45.53	58	70.26
4	5	10.10	14	11.07	25	18.43
5	2	2.47	1	3.00	5	4.65
6	-	-	-	-	1	1.21
	$a = 0.4439$; $\alpha = 0.1062$; $X_2^2 = 3.91$; $P = 0.14$.		$a = 0.4446$; $\alpha = 0.1352$; $X_2^2 = 2.15$; $P = 0.34$.		$a = 0.4646$; $\alpha = 0.1769$; $X_3^2 = 4.60$; $P = 0.20$.	

Text 16: André Gide: Brief an Jacques-Emile Blanche, 1.10..1903 (aus: Collet, Georges-Paul [Hg.], *Correspondance André Gide - Jacques-Emile Blanche: 1892-1939*. [Cahiers André Gide, Bd. 8] Paris, Gallimard, 1979).

Text 17: André Gide: Brief an Jacques-Emile Blanche, 23.10.1918 (Textquelle wie Text 16).

Text 18: André Gide: Brief an Jacques-Emile Blanche, 23.10.1918 (Textquelle wie Text 16).

	Text 19		Text 20		Text 21	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	387	386.98	321	316.16	324	323.86
2	144	144.69	141	142.53	126	126.06
3	40	38.94	44	51.73	59	59.21
4	8	7.82	20	13.66	23	22.87
5	1	1.55	2	3.92	-	-
	$a = 0.4047$; $\alpha = 0.0760$; $X_2^2 = 0.23$; $P = 0.89$.		$a = 0.5128$; $\alpha = 0.1209$; $X_1^2 = 2.37$; $P = 0.12$.		$a = 0.4963$; $\alpha = 0.2157$; $X_1^2 = 0.002$; $P = 0.97$.	

Text 19: Dorothy Bussy: Brief an André Gide, 08.07.1942 (aus: Lambert, Jean [Hg.], *Correspondance André Gide - Dorothy Bussy* [Janvier 1937-Janvier 1951] [Cahiers André Gide, Bd. 11], Paris, Gallimard, 1982).

Text 20: Dorothy Bussy: Brief an André Gide, 14.12.1944 (Textquelle wie Text 19).

Text 21: Dorothy Bussy: Brief an André Gide, 20.5.1946 (Textquelle wie Text 19).

	Text 22	
x	n_x	NP_x
1	254	254.60
2	97	97.63
3	40	37.37
4	9	9.54
5	2	2.86
	$a = 0.4567$; $\alpha = 0.1604$; $X_2^2 = 0.47$; $P = 0.79$.	

Text 22: Dorothy Bussy: Brief an André Gide, 18.11.1949 (Textquelle wie Text 19).

3. Als Ergebnis dieser Untersuchung kann festgestellt werden, daß die Anpassung der Hirata-Poisson-Verteilung an die Brief- und Pressetexte in allen Fällen gelungen ist: Alle Texte erfüllen das Kriterium $P(X^2) \geq 0{,}05$. Da Briefe und Pressetexte recht unterschiedlichen Funktionalstilen angehören, deutet sich an,

daß die gefundene Verteilung ein valides Modell für ein breites Textspektrum des modernen Französisch zu sein scheint. Ob es sich für alle Textsorten auch evtl. für ältere Texte bewährt, kann nur mit weiteren Untersuchungen geklärt werden.

Literatur

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In P. Schmidt (Hg.), *Glottometrika 15* (S. 147-157), Trier: WVT.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In diesem Band,

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In P. Schmidt (Hg.), *Glottometrika 15* (S. 158-165), Trier: WVT.

Grotjahn, R. (1982). Ein statistisches Modell für die Verteilung der Satzlänge. *Zeitschrift für Sprachwissenschaft*, 1, 44-75.

Nemcová, E., & Altmann, G. (1994). The Theory of Word Length: Some Results and Generalizations. *Zeitschrift für empirische Textforschung*, 1, 40-43.

Wimmer, G., & Altmann, G. (1996). The theory of word length: some results and generalizations. In P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Nachschlagewerke:

Le Petite Larousse Illustré 1992. Hg. Paul Robert, Paris: Larousse 1991.

Le Micro Robert - Dictionnaire du Français Primordial. Paris: S.N.L. - Le Robert 1988.

Le Petit Robert. Dictionnaire alphabétique et analogique de la langue française. Paris: Dictionnaires Le Robert 1990.

Glottometrika 16, 1997, 152-162

Untersuchung zur Häufigkeit von Wortlängen in russischen Texten

Patricia Girzig

0. Die bisherigen Untersuchungen zur Wortlängenhäufigkeit in ca. 30 verschiedenen Sprachen (vgl. Best & Zhu, 1994; Uhlířova, 1996; sowie Beiträge in Glottometrika 15 und in diesem Band) und unterschiedlichen Funktionalstilen haben bei aller Verschiedenheit ein gemeinsames Ergebnis: Immer ließen sich Modelle finden, die an die jeweilige Textgruppe angepaßt werden konnten, und es sind, wenn überhaupt, immer nur einzelne Texte, die sich dem jeweiligen Modell entziehen. Die Verschiedenartigkeit der benutzten Modelle läßt sich als spezifische Ausprägung ein und derselben zugrundeliegenden Gesetzlichkeit interpretieren (vgl. Wimmer u.a.., 1994).

In dieser Arbeit geht es darum, an Beispielen russischer Lyrik und Kurzgeschichten darzustellen, wie die Verteilung der Wortlängenhäufigkeit ausfällt und ob sich die erweiterte positive Binominalverteilung, die sich im Tschechischen und Polnischen[1] bereits bewährt hat, auch hier als bestes Modell erweist.

Für diese Untersuchung wurden 31 russische Gedichte vom Anfang des 18. Jahrhunderts bis zur Gegenwart und sieben Kurzgeschichten neuerer sowjetischer Schriftsteller ausgewählt. Die lyrischen Textbeispiele sind chronologisch angeordnet. Von den wichtigsten Literaten des jeweiligen Zeitraums sind mindestens drei Werke aus unterschiedlichen Lebensphasen gewählt worden, um zu prüfen, ob sich ein zeitlicher und sprachlicher Unterschied in den Werken der verschiedenen Autoren in den Daten bemerkbar macht. Die sowjetischen Kurzgeschichten sind in russischer Sprache geschrieben, haben aber neben russischen auch einen kirgisischen (Text 34) und tschuktschischen (Text 35) Verfasser.

Die gewählten Texte sind dem Funktionalstil der schönen Literatur zuzuordnen (vgl. Spillner, 1974:56ff) und die Wahl von zwei Gattungen ermöglicht den

[1] Zum Polnischen haben J. Sambor (Warschau) und W. Lehfeldt (Göttingen) insgesamt 18 Prosatexte ausgewertet, die diesem Modell entsprechen. Zum Tschechischen vgl. L. Uhlířova (1995) und (1996). Im Slowakischen hingegen mußten für die drei bearbeiteten Textklassen (poetische Texte, literarische Prosa und journalistische Texte) andere Modelle gefunden werden. Vgl. Nemcová & Altmann (1994).

direkten Vergleich von kurzen (ab 34 Wörter) und mittleren (bis 1394 Wörter) Textlängen und ihrer Modellierbarkeit.

Die Gedichte wie auch die Kurzgeschichten liegen unter dem von Hammerl (1989:155) und anderen festgestellten kritischen Wert von 2000 Wörtern, um die Homogenität zu wahren.

Als Untersuchungseinheit wurde das graphematische Wort bestimmt, d.h. das durch Leerstellen und Interpunktionszeichen im Schriftbild isolierte Wort (vgl. Bünting & Bergenholz, 1989:36f). Der Bindestrich wird nicht als worttrennendes Interpunktionszeichen gewertet. Im Russischen sind Bindestrichformen auch bei Nichtkomposita, z. B. obligatorisch bei Indefinitpronomen[2], anzutreffen. Die Wortlängen wurden durch Silben, die nach der Zahl der im Wort vorkommenden Vokale und Diphthonge bestimmt sind, gemessen.

Im Russischen sind alle Diphthonge mit *j* zusammengesetzt und treten nicht wie im Deutschen als untrennbare Verbindung von zwei Kurzvokalen auf. Eine Ausnahme bildet der Diphthong *au* in Fremdwörtern (z. B. *šlagbaum*, zweisilbig, Text 20), in echt russischen Wörtern wird er getrennt gesprochen. Dasselbe gilt für *ai*, zum Beispiel: *na-úka* (Wissenschaft), *Ukra-ína* (Ukraine).

Bei den Verben der Bewegung mit Präfix wird der Diphthong Vokal+*j*, obwohl in der Lautung in Richtung Vokal+*i* verschoben und dann getrennt gesprochen, einsilbig gewertet. (z. B. *sojdet*, zweisilbig, Text 4).

Die Entscheidung für das graphematische Wort als Untersuchungseinheit und die Wortlängenmessung in Silben hatte neben der leichteren Identifizierbarkeit und Handhabung in der Untersuchung für die russische Textauswertung besondere Bedeutung:

Im Russischen sind silbenlose, das heißt nur aus Konsonanten bestehende Wörter vertreten. Die historisch gesehen ursprünglich einsilbigen Wörter, hier die aus einem Konsonanten bestehenden Präpositionen *s, k, v*, und die nullsilbigen Varianten einsilbiger Wörter, hier die Partikelvarianten *b, ž, l'*[3], gehen als nullsilbige Wörter in die Dateien ein, daraus ergibt sich die Erweiterung auf die Klasse $x = 0$.

Abkürzungen wurden in ihrer mündlich realisierten Form in den Bestand aufgenommen (z. B. *VDNCH*, viersilbig, Text 34). Französische und lateinische Wort- oder Satzeinschübe sind gemäß ihrer Aussprache bzw. Schreibweise gewertet worden, so z. B. der lateinische Ausruf „In vino veritas!" (Text 20) als drei Wörter von ein, zwei und drei Silben. Der französische Ausdruck *c'est* ins Russische transkribiert zu *se*, wurde als ein einsilbiges Wort bewertet (Text 1).

[2] Die indefiniten Pronomen werden aus Interrogativpronomen mit Partikeln durch Bindestrich verbunden gebildet: *koe-kto* (einige), *kto-to* (jemand), *kto-nibud'* (irgendjemand), *kto-libo* (irgendeiner).

[3] Die Partikel *že* und *by* treten nach Wörtern mit vokalischem Auslaut als nullsilbige Wörter *ž* und *b* auf. Als freie Varianten zählen die Fragepartikel *li* und *l'*. Vgl. Texte 1, 2, 3, 4, 7, 11, 19, 23.

Onomatopoetica, die als Bindestrichformen auftraten, wurden als ein Wort gewertet, z. B. *tik-tak*, zweisilbig, Text 15. Dasselbe gilt für Interjektionen, wie *oj-oj-oj*, die als ein Wort, hier von drei Silben, (Diphthonge), Text 37, gewertet wurden.

2. Die Formel der erweiterten positiven Binominalverteilung lautet:

$$P_x = \begin{cases} 1-a, & x = 0 \\ \dfrac{\alpha \binom{n}{x} p^x q^{n-x}}{1-q^n}, & x = 1, 2, \ldots, n \end{cases}$$

Dabei ist α ein Festparameter, geschätzt als $\alpha = 1 - n_0/N$. Bei der Anpassung der erweiterten positiven Binominalverteilung an Text 8 wurde der Parameter n als bekannt angenommen und mit $n = 6$ angesetzt. In den Tabellen sind angegeben: x (Wortlängenklasse), n_x (Zahl der Wörter mit entsprechender Silbenzahl), NP_x (nach der erweiterten positiven Binominalverteilung berechnete Werte), X^2 (Chiquadrat), P (Überschreitungswahrscheinlichkeit des entsprechenden Chiquadrats).

Die Anpassung wird als zufriedenstellend gewertet, wenn $P \geq 0.05$; Werte von $0.01 \leq P < 0.05$ gelten als noch akzeptabel. Bei schlechtem P-Wert wird auch der Diskrepanzkoeffizient C berücksichtigt, der eine gute Anpassung des Modells an die Daten signalisiert, wenn $C \leq 0.02$.

Die Ergebnisse stellen sich wie folgt dar:

	Text 1		Text 2		Text 3	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	18	18.00	4	4.00	7	7.00
1	91	98.21	57	53.25	44	44.59
2	98	85.62	58	66.21	47	44.75
3	35	37.32	46	41.16	20	22.46
4	5	8.13	13	12.80	7	6.21
5	1	0.72	1	1.61	-	-
	$n = 5$; $p = 0.3036$; $\alpha = 0.9274$; $X_1^2 = 3.38$; $P = 0.07$.		$n = 5$; $p = 0.3834$; $\alpha = 0.9777$; $X_2^2 = 2.07$; $P = 0.35$.		$n = 5$; $p = 0.3342$; $\alpha = 0.9440$; $X_1^2 = 0.49$; $P = 0.48$.	

Text 1: Michail Lomonosov (1711-1765): „Večernee razmyšlenie o božiem veličestve pri slučae velikago severnago sijanija" („Abendliche Gedanken über die Größe Gottes aus Anlaß eines großen Nordlichtes") 1743

Text 2: Michail Lomonosov: „Utrennee razmyšlenie o božiem veličestve" („Morgendliche Gedanken über die Größe Gottes") 1743

Text 3: Michail Lomonosov : „Poslušajte, prošu, čto staromu slučilos' ..."(„Hört zu, ich bitte euch, was einem Alten widerfuhr ") 1747

	Text 4		Text 5		Text 6	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	9	9.00	3	3.00	5	5.00
1	51	47.93	18	19.29	35	34.26
2	41	47.39	26	26.44	34	35.54
3	32	26.77	26	20.14	21	20.48
4	7	9.45	5	9.20	7	7.08
5	3	2.48	2	2.52	2	1.67
6	-	-	1	0.43	-	-
	$n = 8$; $p = 0.2202$; $\alpha = 0.9371$; $X_2^2 = 2.83$; $P = 0.24$.		$n = 7$; $p = 0.3136$; $\alpha = 0.9630$; $X_2^2 = 3.72$; $P = 0.16$.		$n = 7$; $p = 0.2569$; $\alpha = 0.9519$; $X_2^2 = 0.17$; $P = 0.92$.	

Text 4: Aleksandr Puškin (1799-1837): „Brozu li ja vdol' ulic sumnych" („Ob ich lärmende Straßen entlangschlendere ...") 1829

Text 5: Aleksandr Puškin: „K Čaadaevu" („An Čaadaev") 1818

Text 6: Aleksandr Puškin: „K ***" („An ***") 1825

	Text 7		Text 8		Text 9	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	6	6.00	9	9.00	1	1.00
1	44	37.00	45	53.82	53	54.33
2	31	46.48	72	56.96	46	42.43
3	41	29.19	28	32.15	18	20.80
4	7	10.33	2	10.20	9	7.17
5	-	-	8	1.87	0	1.84
6	-	-	-	-	1	0.46
	$n = 5$; $p = 0.3858$; $\alpha = 0.9535$; $X_1^2 = 12.32$ $P = 0.00$; $C = 0.10$.		$n = 6$; $p = 0.2974$; $\alpha = 0.9451$; $X_1^2 = 6.30$; $P = 0.01$.		$n = 18$; $p = 0.0842$; $\alpha = 0.9922$; $X_1^2 = 1.89$; $P = 0.17$.	

Text 7: Aleksandr Puškin: „Iz Pindemonti" („Aus Pindemonti") 1836
Text 8: Michail Lermontov (1814-1841): „Kazac'ja kolybel'naja pesnja" („Wiegenlied der Kosaken") 1838
Text 9: Michail Lermontov: „Smert'" („Der Tod") 1830

	Text 10		Text 11		Text 12	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	4	4.00	7	7.00	7	7.00
1	18	16.62	54	49.24	57	59.03
2	15	17.28	31	39.97	61	57.28
3	10	9.58	22	16.23	30	29.64
4	4	3.53	2	3.58	6	8.63
5	-	-	-	-	2	1.45
	$n = 6$; $p = 0.2937$; $\alpha = 0.9216$; $X_1^2 = 0.50$; $P = 0.48$.		$n = 5$; $p = 0.2887$; $\alpha = 0.9397$; $X_1^2 = 5.21$; $P = 0.02$.		$n = 6$; $p = 0.2796$; $\alpha = 0.9571$; $X_2^2 = 1.35$; $P = 0.51$.	

Text 10: Michail Lermontov: „Čaša žizni" („Der Kelch des Lebens") 1831
Text 11: Michail Lermontov: „Vychožu odin ja na dorogu ..." („Ich gehe allein auf die Straße hinaus ...") 1841
Text 12: Aleksej Tolstoj (1817-1875): „Volki" („Die Wölfe") 1840er Jahre

	Text 13		Text 14		Text 15	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	1	1.00	4	4.00	7	7.00
1	16	19.71	51	53.58	41	44.77
2	23	18.54	58	52.62	40	38.53
3	10	10.17	23	25.84	22	18.42
4	2	3.59	6	6.34	3	5.28
5	2	1.01	1	0.64	2	1.02
	$n = 9$; $p = 0.1904$; $\alpha = 0.9815$; $X_1^2 = 1.85$; $P = 0.17$.		$n = 5$; $p = 0.3293$; $\alpha = 0.9720$; $X_1^2 = 0.99$; $P = 0.32$.		$n = 7$; $p = 0.2229$; $\alpha = 0.9391$; $X_1^2 = 1.33$; $P = 0.25$.	

Text 13: Aleksej Tolstoj :"Ne veter, veja s vysoty ..." („Nicht der Wind, der von der Höhe wehend ...") 1851/52
Text 14: Aleksej Tostoj: „Sleza drožit v tvoem revnivom vzore ..." („Eine Träne zittert in deinem eifersüchtigen Blick ....") 1858
Text 15: Konstantin Bal'mont (1867-1943): „Dožd'" („Regen") 1901

	Text 16		Text 17		Text 18	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	20	20.00	2	2.00	11	11.00
1	144	134.85	19	17.55	63	60.54
2	122	138.96	20	22.31	68	72.01
3	86	79.55	17	16.21	49	48.95
4	31	27.32	7	7.36	23	20.79
5	4	6.32	3	2.59	5	5.65
6	-	-	-	-	1	1.08
	$n = 7$; $p = 0.2557$; $\alpha = 0.9509$; $X_2^2 = 4.55$; $P = 0.10$.		$n = 8$; $p = 0.2664$; $\alpha = 0.9706$; $X_2^2 = 0.49$; $P = 0.78$.		$n = 8$; $p = 0.2537$; $\alpha = 0.9500$; $X_3^2 = 0.64$; $P = 0.89$.	

Text 16: Konstantin Bal'mont: „Pamjati I. S. Turgeneva" („Erinnerungen an I. S. Turgenev") 1893
Text 17: Konstantin Bal'mont: „Podvodnye rasten'ja" („Unterwasserpflanzen") 1894
Text 18: Aleksandr Blok (1880-1921): „Neznakomka" („Die Unbekannte") 1906

	Text 19		Text 20		Text 21	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	7	7.00	0	1.00	6	6.00
1	37	43.47	13	11.68	65	58.38
2	55	47.58	11	12.56	39	51.57
3	37	32.40	8	6.75	33	28.03
4	11	15.37	2	2.03	12	10.48
5	4	5.38	-	-	2	2.85
6	1	1.44	-	-	1	0.71
7	1	0.39	-	-	-	-
	$n = 16$; $p = 0.1274$; $\alpha = 0.9542$; $X_3^2 = 4.39$; $P = 0.22$.		$n = 5$; $p = 0.3496$; $\alpha = 0.9706$; $X_1^2 = 1.57$; $P = 0.21$.		$n = 14$; $p = 0.1196$; $\alpha = 0.9620$; $X_2^2 = 5.00$; $P = 0.08$	

Text 19: Aleksandr Blok: „Nastignutyj metel'ju" („Vom Schneesturm erfaßt") 1907
Text 20: Aleksandr Blok: „Noč', ulica, fonar', apteka ..." („Nacht, Straße, Laterne, Apotheke ...") 1912
Text 21: Aleksandr Blok: „K muze" („An die Muse") 1912

	Text 22		Text 23		Text 24	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	5	5.00	2	2.00	5	5.00
1	24	25.13	18	18.22	37	36.77
2	29	24.93	19	19.65	30	30.14
3	12	15.94	12	12.12	12	12.35
4	8	7.38	5	4.67	3	2.75
5	3	2.64	2	1.37	-	-
6	1	1.00	-	-	-	-
	$n = 31$; $p = 0.0620$; $\alpha = 0.9390$; $X_2^2 = 1.78$; $P = 0.41$.		$n = 8$; $p = 0.2356$; $\alpha = 0.9655$; $X_2^2 = 0.37$; $P = 0.83$.		$n = 5$; $p = 0.2907$; $\alpha = 0.9625$; $X_1^2 = 0.04$; $P = 0.85$.	

Text 22: Anna Achmatova (1889-1966): „Tvorčestvo" („Das Schaffen") 1936
Text 23: Anna Achmatova: „Nam svezest' slov i custva prostotu .." („Bedeutet für uns die Frische der Worte und die Einfachheit des Gefühls ...") 1915
Text 24: Anna Achmatova: „Tebe pokornoj? Ty sošel s uma!" („Dir gehorsam sein? Du bist wahnsinnig!") 1921

	Text 25		Text 26		Text 27	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	6	6.00	14	14.00	7	7.00
1	51	48.35	55	51.41	56	55.29
2	40	45.25	64	73.82	38	40.45
3	26	23.53	64	53.01	22	19.25
4	8	7.34	14	19.03	6	6.70
5	1	1.54	3	2.76	1	1.82
6	-	-	-	-	1	0.51
	$n = 7$; $p = 0.2378$; $\alpha = 0.9545$; $X_2^2 = 1.25$; $P = 0.53$.		$n = 5$; $p = 0.4179$; $\alpha = 0.9346$; $X_2^2 = 5.19$; $P = 0.07$.		$n = 42$; $p = 0.0345$; $\alpha = 0.9466$; $X_2^2 = 0.66$; $P = 0.72$	

Text 25: Anna Achmatova: „Iz vostočnoj tetradi" („Aus dem orientalischen Heft") 1959
Text 26: Boris Pasternak (1890-1960): „V bol'nice" („Im Krankenhaus") 1957
Text 27: Boris Pasternak: „Krasavica moja, vsja stat' ..." ("Meine Schöne, deine ganze Gestalt ...") 1931

	Text 28		Text 29		Text 30	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	4	4.00	4	4.00	2	2.00
1	24	22.91	56	53.50	12	13.66
2	27	27.91	32	36.85	19	16.16
3	14	15.10	16	13.53	8	8.50
4	4	3.08	3	3.13	1	1.68
	$n = 4$; $p = 0.4481$; $\alpha = 0.9452$; $X_1^2 = 0.4462$; $P = 0.50$.		$n = 6$; $p = 0.2160$; $\alpha = 0.9640$; $X_1^2 = 1.21$; $P = 0.27$.		$n = 4$; $p = 0.4410$; $\alpha = 0.9524$; $X_1^2 = 1.00$; $P = 0.32$.	

Text 28: Boris Pasternak: „O, znal by ja, čto tak byvaet .." („Oh, hätte ich gewußt, daß es so ist ...") 1932
Text 29: Bulat Okudžava (*1924): „Odin soldat na svete žil .." („Es lebte einmal ein Soldat ...") 1960
Text 30: Bulat Okudžava: „Pesenka o metro" („Lied über die Metro") 1968

	Text 31		Text 32		Text 33	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	4	4.00	24	24.00	25	25.00
1	29	33.36	262	262.26	235	256.45
2	39	28.90	248	246.39	294	254.54
3	6	11.13	136	140.29	138	149.72
4	1	1.62	59	53.92	54	57.79
5	-	-	13	14.74	11	15.30
6	-	-	2	2.94	5	3.22
7	-	-	1	0.51	-	-
	$n = 4$; $p = 0.3662$; $\alpha = 0.9494$; $X_1^2 = 6.69$; $P = 0.01$.		$n = 12$; $p = 0.1459$; $\alpha = 0.9678$; $X_3^2 = 0.88$; $P = 0.83$.		$n = 10$; $p = 0.1807$; $\alpha = 0.9672$; $X_3^2 = 11.30$; $P = 0.01$; $C = 0.01$.	

Text 31: Bulat Okudžava: „A kak pervaja ljubov' - ona serdce žžet ...“ („Und die erste Liebe, sie verbrennt das Herz ...“)
Text 32: Viktor Dragunskij (1913-1972): „Drug detstva“ („Der Freund aus der Kindheit“)
Text 33: Andrej Platonov (1899-1951): „Neizvestnyj cvetok“ („Die unbekannte Blume“)

	Text 34		Text 35		Text 36	
x	n_x	Np_x	n_x	Np_x	n_x	NP_x
0	45	45.00	58	58.00	45	45.00
1	344	363.12	454	463.68	295	283.91
2	431	402.65	470	450.43	361	375.91
3	259	255.14	262	269.27	283	284.41
4	86	101.04	108	110.67	140	134.48
5	28	25.61	37	33.08	39	40.70
6	3	4.06	3	7.42	8	7.69
7	1	0.41	1	1.27	2	0.90
8	-	-	1	0.22	-	-
	$n = 8$; $p = 0.2406$; $\alpha = 0.9624$; $X_3^2 = 5.57$; $P = 0.13$.		$n = 14$; $p = 0.1300$; $\alpha = 0.9584$; $X_4^2 = 4.62$; $P = 0.33$.		$n = 8$; $p = 0.2745$; $\alpha = 0.9616$; $X_3^2 = 1.57$; $P = 0.67$.	

Text 34: Vasilij Šukšin (1929-1974): „I razygralis' že koni v pole“ („Und die Pferde tollten im Spiel auf dem Feld“)
Text 35: Čingiz Ajtmatov (*1928): „Soldatënok“ („Der kleine Soldat“)
Text 36: Jurij Rytchéu (*1930): „Parusa" ("Die Segel“)

	Text 37		Text 38	
x	n_x	Np_x	n_x	Np_x
0	16	16.00	34	34.00
1	110	118.14	129	146.40
2	138	120.27	214	196.09
3	77	77.55	182	162.59
4	22	35.53	78	93.33
5	17	12.30	35	39.29
6	4	4.25	10	12.53
7	-	-	6	3.81
	$n = 21$; $p = 0.0924$; $\alpha = 0.9583$; $X_3^2 = 10.14$; $P = 0.02$.		$n = 15$; $p = 0.1606$; $\alpha = 0.9506$; $X_4^2 = 10.84$; $P = 0.03$; $C = 0.02$.	

Text 37: Michail Prišvin (1873-1954): „Pikovaja dama“ („Pique Dame“)
Text 38: Ivan Sokolov-Mikitov (1892-1975): „Dorogi“ („Wege“)

3. Die Untersuchung hat gezeigt, daß die Anpassung der erweiterten positiven Binominalverteilung an die lyrischen Texte und die Kurzgeschichten in fast allen Fällen gelungen ist. Ganz inakzeptabel ist lediglich die Anpassung an den Text 7, ein Gedicht von Puškin. Mit einer Modifikation des Modells ließe sich vermutlich auch diese Datei erfassen. Erwähnenswert ist, daß die eingangs erwähnten, nicht zufriedenstellenden, aber noch akzeptablen Werte von $0.01 \leq P < 0.05$ sowohl in den kurzen als auch in den langen Texten, bei verschiedenen Autoren und in unterschiedlichen Entstehungszeiten vorkommen (vgl. Text 8, 11, 31, 33, 37, 38). Im Fall der Texte 33 und 38 zeigen die C-Werte neben den nicht ganz zufriedenstellenden P-Werten, daß die Anpassungen als einigermaßen geglückt gelten dürfen.

Im Unterschied hierzu erwiesen sich in der Untersuchung von 52 Briefen A. S. Puškins (vgl. Stitz, 1994) die erweiterte positive Poissonverteilung, die Hyperpoissonverteilung, in einem Einzelfall die positive Poissonverteilung und die 1-verschobene Poissonverteilung als geeignete Modelle. Das bedeutet, daß sich neben den sprachspezifischen Bedingungen auch historische, autorenspezifische und textsortenspezifische Einflüsse bemerkbar machen.

Es ist eine der Aufgaben zukünftiger Forschung, diesen verschiedenartigen Einflüssen nachzugehen. Als weitere Forschungsperspektive stellt sich die Aufgabe, die Wortlängenverteilungen innerhalb der slawischen Sprachen und ihrer Untergruppen in ihren historischen Zusammenhängen zu untersuchen.

Quellen

Borowsky, K., & Müller, L. (Hg.) (1991). *Russische Lyrik. Von den Anfängen bis zur Gegenwart*. Stuttgart: Reclam.
Gor'kij, M. (Hg.) (1954). *M. V. Lomonosov*. Stichotvorenija. Leningrad.
Gor'kij, M. (Hg.) (1969). *K. D. Bal'mont*. Stichotvorenija. Leningrad.
Dorogi. Rasskazy sovetskich pisatelej. (1988) 3. Aufl. Moskau: Russkij Jazyk.

Literatur

Best, K.-H.,& Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.
Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax*. Grundbegriffe zum Lesen einer Grammatik. 2. überarb. Auflage. Stuttgart: Athenäum.
Hammerl, R. (1990). Untersuchung zur Verteilung der Wortarten im Text. In L. Hřebíček (Hg.), *Glottometrika 11* (S. 142-156), Bochum: Brockmeyer.
Nemcová, E. ,& Altmann, G. (1994). Zur Wortlänge in slowakischen Texten. *Zeitschrift für empirische Textforschung*, 1, 40-43.
Spillner, B. (1974). *Linguistik und Literaturwissenschaft. Stilforschung, Rhetorik, Textlinguistik*. Stuttgart: Kohlhammer.
Stitz, K. (1994). Untersuchung zu den Wortlängen in deutschen und russischen Briefen des 19. Jahrhunderts. Hausarbeit im Rahmen der Ersten Staatsprüfung für das Lehramt an Gymnasien. Göttingen.
Uhlířová, L. (1995). O jednom modelu rozložení délky slov. *Slovo a slovenost* 56, 8-14.
Uhlířová, L. (1996). How long are words in Czech? In P. Schmidt (Hg.), *Glottometrika 15* (S. 134-146), Trier: WVT.
Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Glottometrika 16, 1997, 163-173

Word Length Distribution in Czech: On the Generality of Linguistic Laws and Individuality of Texts

Ludmila Uhlířová

This paper is a free continuation of a paper published in *Glottometrika 15* (1996) in which empirical data on word length in Czech were analysed and a model of statistical distribution of word length for Czech was proposed. The present paper and the former one share a methodological and theoretical functional framework accepted by the participants of a multilingual project on word-length theory started by K.-H. Best (see Altmann & Best, 1996) in Göttingen and outlined in papers by Wimmer & Altmann (1996), Altmann, Erat & Hřebíček (1996), Wimmer, Köhler, Grotjahn & Altmann (1994), as well as in some papers in this volume.

The aim of this paper is to further contribute to the empirical verification of general laws of statistical distribution obeyed by the length of linguistic units. For this purpose, ten short stories for children written by a contemporary Czech writer, M. Macourek, (1991) are analysed, and the statistical distributions of word length in each text are tested and commented on. Commenting on the data, the author adheres to the opinion that, on the one hand, a text, if taken as a whole, obeys very general laws of word-length distribution (the adequacy of which can be tested statistically), and on the other hand, any text represents an individual creative linguistic act, and as such, it is "a population of its own" (to paraphrase Altmann 1993:66; similarly, Hřebíček & Altmann, 1993). In this connection, a question arises how the individuality of a text, resulting from a very complex interplay of a great number of factors of various kinds and strength, is mapped onto its general statistical structure. In other words: Which of the "surface" quantitative "regularities" of word length in texts are the most salient and systematic, and therefore capable of suggesting us something about the expected "deep" probabilistic models which underlie the functioning of human language?

In the following, some such potential "surface" quantitative features are searched for, namely some lexical properties of word classes of different lengths,

the relative repetition rate of f_x in the same text and in different texts and the lower or higher stability of f_x across the texts.

Before presenting the data let us briefly remember that in accordance with the project mentioned above, the word, i. e. the measured unit, is defined at the text level as any *word form* occurring in a running text, and not in its canonic form, as given in dictionaries. Thus, the word-length distribution is always studied in a smaller or wider class of texts, and the results are interpreted with regard to that class. The length of word forms is measured in number of *syllables*. In Czech, any syllabic boundary either falls inside the word form, or coincides with the word form boundary (see Ludvíková, 1985). The only exception is four zero-syllable prepositions *v, k, s, z*, 'in', 'to', 'with' and 'from'; these prepositions, if pronounced (but not in spelling) join the first syllable of the immediately following word, forming a special joint syllable together with it (see Uhlířová 1996 for details and for examples). As our analysis of word length starts from the lexical level (and not from the phonetic level of the syllable), the zero-syllable prepositions are treated as a separate *closed* class of four zero-syllabic words.

The analysed texts are approximately of the same length, about 1000 word forms each, the shortest one consisting of 789 word forms, and the longest one, of 1320 word forms. The writer is a good story-teller. He prefers short clauses with simple syntactic structure, which follow one after another often without any overt syntactic or lexical marker of the semantic connection between them. Syntactic and lexical colloquialisms are frequent. Long and heavy words which may be difficult to understand for children are avoided. All short stories are linked together by the main character, a monkey called *Žofka*, and by the place - a zoological garden and its - personalised - animal inhabitants. Word length as an *intentional* stylistic means is rarely evidenced in Macourek's prosaic, non-rhythmicised texts; perhaps the only cases are several onomatopoetic proper names by which the author characterises positive or negative habits of some of the heroes.

As the statistical computation has shown, the word length distribution in all texts can be modelled by means of the *extended positive binomial distribution*, with one *local modification* occurring in two texts (on the notion of local modification see Altmann, 1993). If x is a variable denoting word length in a text consisting of N running words, and if P_x is the probability that a word form will be x syllables long (= theoretical frequency of the class of word forms of length x), and if $P_0 = 1 - \alpha$, then

$$P_x = \frac{\alpha \binom{n}{x} p^x q^{n-x}}{1 - q^n}$$

for all $x = 1, 2, ..., n$.

As can be seen from Table 1, this type of distribution is a model which fits best for all of Macourek's texts with the exception of texts number 2 and number 4.

Table 1:

	Text 1		Text 2		Text 3	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	18	18.00	28	28.00	26	26.00
1	371	369.56	448	455.26	465	460.93
2	392	391.75	443	419.57	398	407.42
3	218	221.47	183	206.23	215	205.78
4	71	70.43	67	57.02	64	64.96
5	13	11.95	5	8.41	9	13.12
6	1	0.86	1	0.54	3	1.80
	$n = 6$; $p = 0.29776$; $\alpha = 0.98339$; $X_2^2 = 0.18$; $P = 0.91$.		$n = 6$; $p = 0.26935$; $\alpha = 0.97617$; $X_2^2 = 6.75$; $P = 0.03$; $C = 0.006$.		$n = 8$; $p = 0.20162$; $\alpha = 0.97617$; $X_3^2 = 2.81$; $P = 0.42$.	

	Text 4		Text 5		Text 6	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	27	27.00	33	33.00	28	28.00
1	397	419.00	414	438.59	295	293.98
2	420	369.35	388	388.20	274	279.04
3	165	197.32	199	190.89	148	141.26
4	78	71.15	69	56.32	38	40.22
5	14	18.25	14	9.97	4	6.11
6	5	3.96	1	1.04	2	0.41
	$n = 12$; $p = 0.13813$; $\alpha = 0.97559$; $X_3^2 = 15.33$; $P = 0.002$; $C = 0.014$.		$n = 7$; $p = 0.22782$; $\alpha = 0.97048$; $X_3^2 = 6.20$; $P = 0.10$.		$n = 6$; $p = 0.27519$; $\alpha = 0.96451$; $X_2^2 = 0.58$; $P = 0.75$.	

	Text 7		Text 8		Text 9	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	34	34.00	45	45.00	37	37.00
1	483	475.28	469	492.43	418	423.53
2	449	470.34	479	445.92	394	376.37
3	268	248.24	240	235.55	161	178.38
4	69	73.70	72	79.99	56	47.56
5	10	11.67	11	18.11	4	7.18
6	1	0.79	2	2.73	-	-
7	-	-	2	0.32	-	-
	$n = 6$; $p = 0.28359$; $\alpha = 0.97412$; $X_2^2 = 3.13$; $P = 0.21$.		$n = 9$; $p = 0.18460$; $\alpha = 0.96591$; $X_3^2 = 7.56$; $P = 0.06$.		$n = 6$; $p = 0.26224$; $\alpha = 0.96542$; $X_3^2 = 5.49$; $P = 0.06$.	

	Text 10	
x	f_x	NP_x
0	34	34.00
1	413	424.07
2	381	371.36
3	183	180.66
4	51	52.74
5	9	9.24
6	2	0.96
	$n = 7$; $p = 0.22594$; $\alpha = 0.96831$; $X_2^2 = 0.69$; $P = 0.71$.	

The latter two texts can be appropriately modelled by the extended positive binomial distribution only on the assumption that the word classes of lengths $x = 2$ and $x = 3$ are given a special treatment. The word class of length $x = 2$ seems to be foregrounded in the structures of these texts, whereas the word class of length $x = 3$ is backgrounded.

Therefore, if

$$P_x' = \begin{cases} 1-\alpha & x=0 \\ \alpha P_1 & x=1 \\ \alpha P_2(1+bP_3) & x=2 \\ \alpha P_3(1-b) & x=3 \\ \alpha P_x & x=4,\ldots,n \end{cases}$$

then we get a local modification of the extended positive binomial distribution; the results of the χ^2-test are very good for both cases (see Table 1).

Table 2: (the fitting obtained for Texts 2 and 4 when equation (2) is applied)

	Text 2		Text 4	
x	f_x	NP_x	f_x	NP_x
0	28	28.00	27	27.00
1	448	445.08	397	419.00
2	443	446.64	420	369.35
3	183	185.61	165	197.32
4	67	60.03	78	71.15
5	5	9.07	14	18.25
6	1	0.57	5	3.96
	$n = 6$; $p = 0.2742$; $\alpha = 0.9762$; $b = 0.1237$; $X_1^2 = 2.27$; $P = 0.13$.		$n = 12$; $p = 0.1424$; $\alpha = 0.9765$; $b = 0.1969$ $X_2^2 = 2.43$; $P = 0.30$.	

Generally, it holds that frequencies of word forms of different lengths are proportional to each other in continuous texts. However, due to the presence of various *boundary conditions* in some of the texts (for the notion of boundary condition see Altmann, 1993) the proportionality should not be expected to realize itself in the same way in any text.

Firstly, it does not hold for all of the analysed texts and for all of the word lengths that if the word length increases, the frequency of words of a given length decreases. In three texts (texts number 1, 4 and 8) f_2 is higher than f_1, and in another one (text number 2) both values are almost equal. In the six remaining texts f_1 is higher than f_2, but the difference between f_1 and f_2 is quite small, ranging from about 1% to 5%. It is important that a high f_2 is not at the expense of a low f_1, but clearly at the expense of a low f_3. The drop of f_3 is biggest in texts number 2 and 4, i.e. in texts with locally modified statistical distributions, as described above.

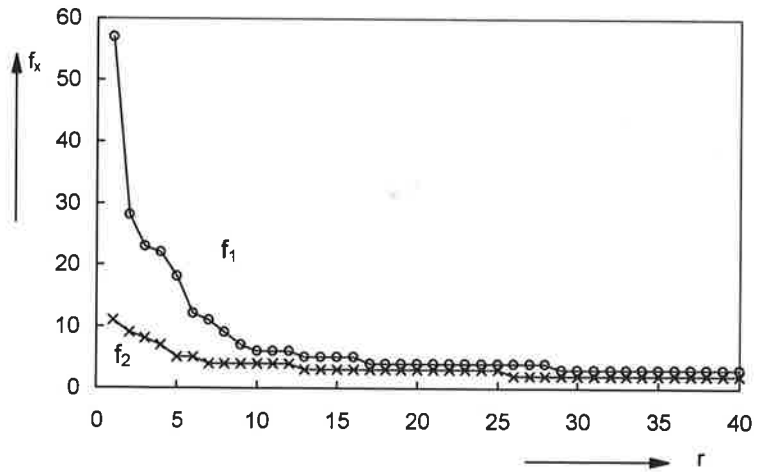
Secondly, it is worth noting that different classes of f_x consist of words of different lexical properties. Most monosyllabic words are grammatical, the conjunction *a* 'and' being the most frequent in each text; it is followed by present tense forms of 'be', such as *je, jsou, jsem*, by the reflexive pronoun *se/si* ('-self'), by prepositions, such as *na, se, za, do* ('in', 'with', 'behind', 'into'), the demonstrative *to* ('this/that'), the relative *co* ('which/what'), as well as by conjunctions, such as *když, že, jak* ('when','that', 'how'); the ten most frequent word forms are not identical in all texts, but their intersection as well as the coincidence of their respective ranks are highly significant. If the class f_0 is left out of consideration for the moment (it is closed and small and as such it is hardly comparable with the other, open, classes), then the word class of f_1 has the highest repetition rate, measured as a ratio of the sum of different word forms of a given class to their total frequency in a text); the repetition rate lies in the interval <4.19 - 3,39>. In other words, this word class consists of a relatively small number of word forms which frequently occur in text(s), and which are highly polysemic, some of them even homonymous. The ten most frequent words of the class of f_1 cover one fifth of each text, which may be considered to be a specific characteristic of the given texts (if the length of the text increases, the value of this characteristic decreases).

In contrast to f_1, all of the other word classes with $f_{x>1}$ show low repetition rates, reaching a maximal value of 1.6 and decreasing with longer word classes; all of these word classes consist of a relatively great number of word forms, which, however, do not occur very often in texts. In the word class of f_2, both words with purely grammatical meanings and words with full lexical meanings occur; the co-occurrence of the same word forms in different texts is not so high as it is in the class of f_1 (see above), but the lexical *categories* of words are almost the same: They are two-syllable (proper) names of the main heroes, connective expressions (including some which clearly mark the author's individual style, e.g. a subordinate conjunction *načež* 'then'), some verb forms of *říkat* 'dicere' and *být* 'esse' (e.g., negative forms *není, nejsou* for 3rd person sg. and pl. present tense, as well as a future tense form *bude* for 3rd person). On the contrary, two-syllable names of inanimate objects, names of places and time occur rarely, and an adjective denoting a quality occurred just once in this class.

Word forms in the class of f_3 almost exclusively belong to the thematic layer of vocabulary. They represent 15 - 20% of all word forms in texts belonging to a great number of various semantic classes, and their rate of repetition is lower than that with the class of f_2. And finally, the word classes of f_4 and longer together make up no more than 6 - 7% of the whole texts; in some cases, their rate of repetition equals 1; in other words, all word forms of a given length occur just once in the respective text. Word forms six and seven syllables long occur rarely, only in some texts, and word forms longer than seven syllables do not occur at all in the corpus.

Let us return to the classes of f_1 and f_2 once more. Although, as stated above, their frequencies are quite near to each other in the analysed texts, there is a substantial difference between the two classes: They represent words of different categorial meanings, which play different roles in texture and, which, as such, also have different rates of repetition. If we make a graph of f_1 and of the respective ranks r in a text (the so-called "iota-curve", see Hammerl & Sambor, 1993), and if we compare it with a graph of f_2 and of the respective ranks in the same text, we obtain two significantly different regression functions with different parameters.

As an illustration, data from text number 1 is presented in Table 3 and the frequency-rank plot is drawn below.



The ranks of words with frequencies f_1 and f_2, cf. Table 3

Table 3:

	Class f_1		Class f_2	
Rank	Word form	Frequency	Word form	Frequency
1.	a	57	Žofka	11
2.	se	28	paní	9
3.	to	23	ale	8
4.	je	22	řekne	7
5.	na	18	ani	5
6.	si	12	není	5
7.	tak	11	bude	4
8.	co	9	načež	4
9.	pan	7	tohle	4
10.	nám	6	všecko	4
11.	jsou	6	kohout	4
12.	chce	6	jako	4
13.	ta	5	jednou	3
14.	že	5	ráno	3
15.	za	5	Milly	3
16.	už	5	celá	3
17.	jak	4	zvláštní	3
18.	před	4	nějak	3
19.	ten	4	Žofko	3
20.	já	4	zrovna	3
21.	kdo	4	volá	3
22.	mít	4	říká	3
23.	nás	4	budí	3
24.	i	4	prosím	3
25.	tu	4	jasné	3
26.	když	4	párek	2
27.	jde	4	nikdo	2
28.	do	4	napůl	2
29.	ty	3	moje	2
30.	dva	3	žena	2
31.	má	3	ovšem	2
32.	dvě	3	která	2
33.	u	3	vlastně	2
34.	mi	3	Willy	2
35.	den	3	nebo	2
36.	ve	3	všechno	2
37.	byt	3	mají	2
38.	jsme	3	tedy	2
39.	jsem	3	kdekdo	2
40.	pro	3	vidět	2
...	...	...	...	...
Total		97		271

Naturally, one cannot expect that even grammatical words are used evenly in texts. The frequencies of grammatical words are text-conditioned as well. This may be easily demonstrated by comparing the variability of frequency classes f_1, f_2 and f_3, i.e. the variability of classes with different proportion of grammatical words, in our texts. The relative frequencies of the same class vary across the texts, as shown by the data below (in %):

interval of minimal to maximal f_1 < 34.2% - 39.4%>
interval of minimal to maximal f_2 < 33.7% - 37.7%>
interval of minimal to maximal f_3 < 14.9% - 20.4%>

Thus, e.g., the lowest relative frequency of monosyllabic word forms (i.e. the class consisting mostly of grammatical words) is 34.2% (in text number 1), whereas the highest relative frequency of monosyllabic word forms is 39.4% (in text number 9). The frequency stability across the texts is highest with disyllabic word forms; the range of frequencies is only 4%, i.e., it is smaller than with monosyllabic word forms (5.2%), and smaller than with trisyllabic word forms (5.3%).

A high relative stability across the texts is characteristic also for tetrasyllabic word forms as well as for all word forms longer than four syllables. This fully confirms an earlier empirical observation based on other Czech data (Uhlířová, 1995): An analysis of ten short stories by a contemporary Czech prosaist, Hrabal, also showed the relatively highest frequency stability of disyllabic word forms across texts; incidentally, in Hrabal's texts (of variable length) the range of relative frequency values was also 4%.

The empirical observations made above lead to the following hypothesis: There are groups of texts - written by the same author and belonging to the same genre - for which it holds that if we want to find specific quantitative characteristics of the author's style, we should concentrate on f_1 and on f_3 in more detail. More generally, if the relationship of proportionality between classes f_x is to be interpreted, and if local modifications of probability values of NP_x are to be interpreted, attention should be paid not only to the proportionality of neighbouring classes, but also to non-neighbouring classes; in some cases, the latter may provide even more information and thus may be more revealing for interpretative purposes.

To sum up: The data presented above from Macourek's ten short stories for children confirm what was stated earlier (Uhlířová, 1996), namely that Czech fits, so far, *one* type of word-length distribution - the extended positive binomial distribution, which, in some texts, is locally modified in various ways. Texts belonging to different genres and written by different authors lead to different modifications. In two of Macourek's texts, the proportion between f_2 and f_3 must have

been paid special attention. On the other hand, in some of Hrabal's short stories (Uhlířová, 1996) it was the word class of length $x = 1$ which seemed to be foregrounded, whereas the word class of length $x = 2$ seemed to be backgrounded. And finally, in some journalistic texts (Uhlířová, 1996) the foregrounded class was $x = 3$ at the expense of $x = 4$. Thus, the original model of the extended binomial distribution becomes more and more diversified if more texts are examined, but, for the time being, it *still remains within the limits of local modifications* of some of the classes.

In all cases studied so far the frequency shifts are directed from longer word forms to shorter ones, and not conversely, thus conforming to the general relationships described by linguistic laws (such as Menzerath's law, see Hřebíček & Altmann, 1993) as well as to the general linguistic knowledge of the development of human language towards a truly economical and well-balanced communicative system. The empirical fact that texts of one and the same author are characterised, so far, by one distribution in the default shape plus no more than ONE local modification, strongly supports the basic idea of synergetic functional linguistics, namely that any text is produced - as a new quality - under certain starting conditions and by certain strategies, which operate either in conformity to or against one another, and that the starting conditions remain, at least in unmarked cases, reasonably stable during the process of text production. The starting conditions and strategies seem to preserve their stability more easily if texts of one and the same genre (though not necessarily written by the same author) are produced. On the other hand, different local modifications seem to indicate deeper differences between genres.

References

Altmann, G. (1993). Phoneme counts. Marginal remarks to Pääkkönen's article. In G. Altmann (Ed.), *Glottometrika 14* (pp. 54-68), Trier: WVT.

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 166-180), Trier: WVT.

Altmann, G., Erat, E., & Hřebíček, L. (1996). Word length distribution in Turkish texts. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 195-204), Trier: WVT.

Hammerl, R., & Sambor, J. (1993). *O statystycznych prawach językowych*. Warszawa: Zakład Semiotyki Logicznej UW.

Hřebíček, L., & Altmann, G. (1993). Prospects of text linguistics. In L. Hřebíček & G. Altmann (Eds.), *Quantitative Text Analysis* (pp. 1-28), Trier: WVT.

Ludvíková, M. (1985). Kvantitativní charakteristiky českých fonémů. In M. Těšitelová a kol., *Kvantitativní charakteristiky současné češtiny*, (pp. 11-28), Praha: Academia.

Macourek, M. (1991). *Žofka ředitelkou zoo*. Praha: Mladá fronta.

Uhlířová, L. (1995). O jednom modelu rozložení délky slov [On a model of word-length distribution]. *Slovo a slovesnost*, 56, 8-14.

Uhlířová, L. (1996). How long are words in Czech? In P. Schmidt (Ed.), *Glottometrika 15* (pp. 134-146), Trier: WVT.

Wimmer, G., & Altmann, G. (1996). The theory of word length: Some results and generalizations. In P. Schmdit (Ed.), *Glottometrika 15* (pp. 112-133), Trier: WVT.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

This paper was supported by a grant from GACR 102/96/k087

Glottometrika 16, 1997, 174-179

Wortlängenhäufigkeiten in althebräischen Texten

Claudia Balschun

0. In dieser Arbeit geht es darum, biblische Texte, hier speziell Psalmentexte, daraufhin zu untersuchen, mit welcher Häufigkeit Wörter unterschiedlicher Länge in abgeschlossenen Texten vorkommen. Ziel ist es, nachzuweisen, daß die Wortlängenhäufigkeiten Gesetzen folgen.

Wenn man eine solche Aufgabenstellung verfolgt, ist es wichtig, daß man möglichst homogene Texte zur Verfügung hat. Nun ist bei alten Liedtexten, und um solche handelt es sich ja bei den Psalmen, mit Inhomogenitäten zu rechnen, da die Texte eventuell im Laufe der Zeit verändert worden sein können. Mit Texthomogenität ist am ehesten dann zu rechnen, wenn ein Text von einem Verfasser in einem Zug geschrieben und dann nie wieder verändert wurde. Mit einer solch idealen Entstehungs- und Überlieferungsgeschichte ist im Falle der Psalmen nicht unbedingt zu rechnen. Wenn dennoch die Psalmen als Gegenstand der Untersuchung gewählt wurden, so hat das zwei Gründe:

1. Der Vergleich verschiedener Bibelausgaben zeigt, daß die Kapiteleinteilung nicht immer übereinstimmt; d.h. in solchen Fällen ist unklar, wo die Grenzen eines Textes bzw. Textabschnittes liegen. Tatsächlich haben erste Versuche mit biblischen Prosatexten des Alten Testaments zu erheblichen Problemen geführt.
2. Bisherige Versuche mit lyrischen Texten (vgl. Christiansen, 1997 zu deutschen Barockgedichten und Girzig, 1997 zu russischen Gedichten) haben gezeigt, daß Gedichte offenbar ein gutes Untersuchungsfeld darstellen, obwohl man hier mit vielfachen Korrekturen und Überarbeitungen rechnen muß: Nur für sehr wenige dieser Texte ließ sich kein Modell finden.

Aufgrund dieser Tatsachen bietet es sich an, auch mit althebräischen Liedtexten einen Versuch zu wagen. Bisher wurden insgesamt 23 Psalmen bearbeitet, die relativ willkürlich aus dem Gesamtkorpus ausgewählt wurden.

1. Will man die Wortlängenhäufigkeit von Texten untersuchen, muß man definieren, was als „Wort" gelten soll und wie die Wortlänge zu messen ist. In Übereinstimmung mit anderen Arbeiten (s.o.) wird als „Wort" das „orthographische Wort" bestimmt (Bünting & Bergenholtz, 1989:36); die Wortlänge wird in der Zahl der Silben pro Wort gemessen. Kriterium für „Silbe" ist das Vorkommen von Vokalen oder Diphthongen. Die Untersuchung orientiert sich damit an den in Best & Zhu (1994:20) angegebenen Kriterien.

Hinsichtlich der Anwendung der Kriterien für die Bestimmung eines „Wortes" gibt es im Althebräischen keine Probleme. Die hier sehr häufig vorkommenden Bindestrichverbindungen wurden als ein Wort gewertet. Die Möglichkeit, aus Konsonanten und Vokalen Silben aufzubauen, sind im Althebräischen beschränkter als z.B. im Deutschen. Mit wenigen Ausnahmen gelten die folgenden Regeln (K = Konsonant, V = Vokal):

- Mögliche Silben sind: KV; KVK; KVKK// (// Wortende) (vgl. Jenni, 1981: 38).
- Diphthonge kommen im Hebräischen nicht vor.

Die Psalmen wurden immer vollständig ausgewertet. Sie haben in der „Biblia Hebraica" keine vom eigentlichen Text abgesetzten Überschriften.

2. Die Berechnungen wurden mit Hilfe des Altmann-Fitters (1994) durchgeführt. An alle Texte konnte die 1-verschobene Hyperpoisson-Verteilung angepaßt werden. Die Formel dieser Verteilung lautet:

$$P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1(1;b;a)}, \quad x = 1,2,3,...$$

$$c^{(0)} = 1$$

$$c^{(x)} = c(c+1)...(c+x-1)$$

$${}_1F_1(1;b;a) = \sum_{j=0}^{\infty} \frac{a^j}{b^{(j)}}$$

Die Ergebnisse dieser Anpassung finden sich in den folgenden Tabellen. Dabei bedeutet x die Wortlänge, n_x die beobachtete Zahl der Wörter mit Wortlänge x, NP_x die nach der Hyperpoisson-Verteilung berechnete Zahl der Wörter der Länge x. Ferner ist X^2 das Chiquadrat, P die Wahrscheinlichkeit, den gefundenen Chiquadratwert oder einen höheren zu finden; $C = X^2/N$ ist der Diskrepanzkoeffizient. Die Anpassung gilt als zufriedenstellend, wenn $P \geq 0.05$ oder $C \leq 0.02$. Werte von $0.05 > P \geq 0.01$ gelten als noch akzeptabel. Die Untersuchung hat folgende Ergebnisse erbracht:

	Psalm 5		Psalm 7		Psalm 10	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	8	6.89	3	3.06	6	6.61
2	31	34.90	56	57.12	66	72.66
3	38	32.20	47	42.57	62	46.42
4	13	16.35	12	16.19	10	15.27
5	6	5.72	6	5.07	1	4.05
6	2	1.95	-	-	-	-
	$a = 1.1283$; $b = 0.2228$; $X_3^2 = 2.36$; $P = 0.50$.		$a = 0.7763$; $b = 0.0416$; $X_2^2 = 1.74$; $P = 0.42$.		$a = 0.6783$; $b = 0.0617$; $X_2^2 = 10.00$; $P = 0.01$.	

	Psalm 16		Psalm 18		Psalm 19	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	3	3.39	20	20.69	6	6.08
2	22	24.84	120	124.12	48	47.34
3	28	23.13	119	114.67	37	37.86
4	14	11.49	62	57.39	16	15.96
5	1	5.17	20	19.69	6	5.78
6	-	-	2	6.48	-	-
	$a = 1.0663$; $b = 0.1454$; $X_1^2 = 1.54$; $P = 0.2146$.		$a = 1.0920$; $b = 0.1820$; $X_3^2 = 3.77$; $P = 0.29$.		$a = 0.8913$; $b = 0.1144$; $X_2^2 = 0.04$; $P = 0.98$.	

	Psalm 21		Psalm 22		Psalm 25	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	8	0.58	14	15.03	9	9.27
2	28	34.99	78	83.75	46	47.37
3	33	33.50	82	74.14	45	43.75
4	16	16.16	40	35.65	24	22.21
5	7	6.79	8	11.76	8	7.77
6	-	-	2	3.70	1	2.64
	$a = 0.9728$; $b = 0.0161$; $X_1^2 = 0.02$; $P = 0.88$.		$a = 1.0525$; $b = 0.1890$; $X_3^2 = 3.80$; $P = 0.28$.		$a = 1.1273$; $b = 0.2206$; $X_3^2 = 1.24$; $P = 0.74$.	

	Psalm 26		Psalm 31		Psalm 33	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	3	2.93	14	12.93	12	13.03
2	27	25.64	67	78.10	50	54.28
3	22	25.16	81	66.38	51	43.23
4	15	13.08	31	30.34	19	19.03
5	5	4.62	7	12.27	4	5.79
6	1	1.58	-	-	1	1.66
	$a = 1.1052$; $b = 0.1264$; $X_3^2 = 0.99$; $P = 0.80$.		$a = 0.9892$; $b = 0.1638$; $X_2^2 = 7.15$; $P = 0.03$.		$a = 0.9846$; $b = 0.2363$; $X_3^2 = 2.62$; $P = 0.45$.	

	Psalm 34		Psalm 35		Psalm 38	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	6	5.72	7	7.20	5	6.10
2	55	52.42	80	82.30	43	52.43
3	47	47.94	73	70.34	65	49.20
4	25	23.08	33	31.23	24	24.41
5	2	7.54	9	9.36	6	10.88
6	2	1.86	1	2.59	-	-
	$a = 1.0160$; $b = 0.1108$; $X_3^2 = 5.64$; $P = 0.13$.		$a = 0.9237$; $b = 0.0808$; $X_3^2 = 1.25$; $P = 0.74$.		$a = 1.0532$; $b = 0.1225$; $X_2^2 = 9.15$; $P = 0.01$.	

	Psalm 44		Psalm 49		Psalm 55	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	12	13.31	15	16.34	10	10.91
2	51	56.59	57	62.09	64	69.85
3	63	55.61	55	42.94	68	57.40
4	36	30.90	15	16.33	24	25.20
5	11	16.61	1	5.32	7	9.65
	$a = 1.2783$; $b = 0.3008$; $X_2^2 = 4.39$; $P = 0.11$.		$a = 0.8453$; $b = 0.2224$; $X_1^2 = 5.38$; $P = 0.02$.		$a = 0.9428$; $b = 0.1473$; $X_2^2 = 3.30$; $P = 0.19$.	

	Psalm 57		Psalm 68		Psalm 105	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	4	4.07	26	29.99	9	8.89
2	37	37.69	100	115.34	107	105.69
3	34	32.21	124	91.00	90	87.78
4	13	14.43	32	39.99	36	37.77
5	5	4.38	10	15.68	7	10.97
6	1	1.25	-	-	5	2.91
	$a = 0.9414$; $b = 0.1018$; $X_3^2 = 0.38$; $P = 0.94$.		$a = 0.9925$; $b = 0.2581$; $X_2^2 = 18.18$; $P = 0.0001$; $C = 0.06$.		$a = 0.8929$; $b = 0.0751$; $X_3^2 = 3.12$; $P = 0.37$.	

	Psalm 107		Psalm 145	
x	n_x	NP_x	n_x	NP_x
1	8	8.61	3	2.81
2	89	95.79	40	37.51
3	104	87.87	42	43.67
4	34	42.03	23	26.58
5	12	13.60	16	10.95
6	3	3.32	2	4.50
7	2	0.80	-	-
	$a = 0.9997$; $b = 0.0899$; $X_3^2 = 5.41$; $P = 0.14$.		$a = 1.2755$; $b = 0.0957$; $X_3^2 = 4.42$; $P = 0.22$.	

3. Es läßt sich nun feststellen, daß von den 23 bearbeiteten Psalmen 18 gut mit der 1-verschobenen Hyperpoisson-Verteilung modelliert werden können; in vier Fällen (Psalm 10, 31, 38, 49) konnten nur schwache, wenn auch noch akzeptable P-Werte gefunden werden; in einem Fall (Psalm 68) ließ sich das Modell nicht anpassen. Man kann daher davon ausgehen, daß auch die Wortlängenhäufigkeiten althebräischer Texte sich gesetzmäßig verhalten. Wenn einer unter 23 Texten dem zu widersprechen scheint, kann das vielfältige Ursachen haben, z.B. Störung der Texthomogenität aufgrund von Eingriffen in den ursprünglichen Wortlaut des Textes.

Es hat sich inzwischen allerdings gezeigt, daß die Homogenitätsprobleme bei alten Texten keineswegs so groß sind, daß diese für Wortlängenuntersuchungen deshalb generell nicht in Frage kämen. So ließen sich alle bearbeiteten griechischen Texte (Koine: Bibeltexte, Epiktet) ebenfalls mit der 1-verschobenen Hyperpoisson-Verteilung modellieren (Egbers, Groen, Podehl & Rauhaus, 1997); auch für lateinische Texte (Briefe des Plinius) ließen sich in allen Fällen geeigete Modelle finden, meist die positive Binomialverteilung (Röttger & Schweers, 1997).

Textquelle:

Kittel, R. (Hg.) (1987). *Biblia Hebraica Stuttgartensia*. 3. verb. Auf. Stuttgart: Biblia-Druck.

Literatur:

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax*. Frankfurt: Athenäum.

Christiansen, B. (1997). Wortlängenverteilungen in deutschen Barockgedichten. In diesem Band.

Egbers, J., Groen, C., Podehl, R., & Rauhaus, E. (1997). Zur Wortlängenhäufigkeit in griechischen Koine-Texten. In diesem Band.

Girzig, P. (1997). Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. In diesem Band.

Jenni, E. (1981). *Lehrbuch der hebräischen Sprache des Alten Testaments*. Neubearbeitung des „Hebräischen Schulbuchs" von Hollenberg-Budde. Frankfurt: Helbing & Lichtenhahn.

Röttger, W., & Schweers, A. (1997). Wortlängenhäufigkeit in Plinius-Briefen. In diesem Band.

Software:

Altmann-FITTER (1994), Lüdenscheid: RAM - Verlag.

Glottometrika 16, 1997, 180-184

Wortlängenhäufigkeiten in japanischen Pressetexten

Gesa Riedemann

1. Gegenstand dieser Arbeit ist die Frage, mit welcher Häufigkeit Wörter unterschiedlicher Länge in japanischen Texten vorkommen und ob diese Häufigkeiten einem der bekannten Modelle folgen (vgl. dazu Wimmer & Altmann, 1996). Es handelt sich um insgesamt 11 Texte aus verschiedenen Ausgaben der bekannten Tageszeitung „Asahi Shinbun".

2. Die Untersuchung wurde wie folgt durchgeführt: Als „Wort" wurde das grammatische Wort gewählt. In Übereinstimmung mit Sanada (1993:124f) wurden Fälle wie „Substantiv + suru" oder „Substantiv + na" jeweils als ein Wort gewertet. Die Wortlänge wurde nach der Zahl der Silben bestimmt; die Einteilung der Silben erfolgte nach dem Hiragana- bzw. Katakana-Silbensystem von J.C. Hepburn (Hepburn-Silbensystem: Lewin, 1990). In Zweifelsfällen wurde auf die einschlägige Fachliteratur (z.B. Lewin, 1990) zurückgegriffen.

3. In den folgenden Tabellen werden die Ergebnisse zum Japanischen präsentiert. Die Berechnungen wurden mit dem Altmann-Fitter durchgeführt, Es zeigt sich, daß die 11 Texte alle der 1-verschobenen Hirata-Poisson-Verteilung folgen, die wie folgt lautet:

$$P_x = \sum_{i=0}^{\left[\frac{x-1}{2}\right]} \binom{x-1-i}{i} \frac{e^{-a} a^{x-1-i}}{(x-1-i)!} \alpha^i (1-\alpha)^{x-1-2i}, \quad x = 1, 2, \ldots$$

Zu jedem Text werden folgende Werte angegeben: X ist die Wortlänge, angegeben in der Zahl der Silben des Wortes; n_x ist die festgestellte Anzahl der Wörter des betreffenden Textes mit der entsprechenden Silbenzahl; NP_x gibt die Werte an, die nach der Hirata-Poisson-Verteilung berechnet wurden. X^2 ist das Chiquadrat, P die Überschreitungswahrscheinlichkeit des Chiquadrats; a und α sind Parameter. Das Ergebnis der Anpassung eines Textes an die Hirata-Poisson-Verteilung wird als zufriedenstellend betrachtet, wenn $P \geq 0.05$ oder $C \leq 0.02$.
Die Untersuchung hat folgende Ergebnisse erbracht:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	102	103.66	58	57.72	81	79.95
2	92	92.41	40	42.94	79	77.44
3	65	60.64	30	26.59	40	45.92
4	37	29.58	14	11.86	25	20.26
5	7	12.28	3	4.65	8	7.32
6	0	4.41	1	2.24	1	3.11
7	2	2.02	-	-	-	-
	$a = 1.0792$; $\alpha = 0.1739$; $X_1^2 = 0.448$; $P = 0.5032$.		$a = 0.9278$; $\alpha = 0.1982$; $X_3^2 = 2.284$; $P = 0.5155$.		$a = 1.0738$; $\alpha = 0.0980$; $X_3^2 = 3.397$; $P = 0.3344$.	

Text 1: Die Kaiserin fiel um (Kôgô-sama taoreru). (aus: Asahi Shinbun, Ausgabe Nr.38691, 1993, S.1).
Text 2: Puppen, die Rasenhaare hervorbringen (Shibafu môhatsu haeru ningyô). (aus: Asahi Shinbun, Nr.38861, 1994, S.15. Sparte „Familie").
Text 3: Shaga, die seltene, immergrüne Schwertlilie ist haltbar und auch reich an Wachstumskraft (mezurashii jôryoku no ayame jôbu de hanshoku chikara mo sakan). (aus: Asahi Shinbun, Nr.38861, 1994, S.15. Sparte „Familie").

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	163	168.86	147	145.87	163	162.69
2	89	91.57	93	96.26	84	89.23
3	89	72.86	52	49.63	73	68.47
4	32	30.53	23	18.78	36	28.60
5	9	14.50	4	8.46	9	13.18
6	3	5.04	-	-	3	4.54
7	0	1.83	-	-	0	1.60
8	0	0.55	-	-	1	0.69
9	1	0.26	-	-	-	-
	$a = 0.8267$; $\alpha = 0.3440$; $X_4^2 = 7.833$; $P = 0.0979$.		$a = 0.7824$; $\alpha = 0.1565$; $X_2^2 = 3.516$; $P = 0.1724$		$a = 0.8189$; $\alpha = 0.3302$; $X_4^2 = 5.078$; $P = 0.2794$	

Text 4: Der Polarforscher Akasofu Shunichi (ôrora kenkyusha Akasofu hunichi). (Text aus: Asahi Shinbun, Nr. 38861, 1994, S.3. Sparte „Menschen").
Text 5: Zeit des Menschen. In einer Notiz über großartige Freundlichkeit (Hito toki. Suteki-na shinsetsu nôto ni). (Text aus: Asahi Shinbun, Nr.38861, 1994, S.15. Sparte „Familie").
Text 6: Die Künstlerin Ono Yoko (Achisuto Ono Yoko). (Text aus: Asahi Shinbun, Nr. 38691, 1993, S.3. Sparte „Menschen").

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	68	68.41	88	86.95	77	74.83
2	69	65.16	50	50.29	43	45.69
3	35	38.22	49	50.69	34	36.79
4	21	16.70	39	23.71	22	16.78
5	1	5.98	6	13.96	8	8.17
6	2	1.84	3	9.40	2	3.04
7	1	0.69	-	-	0	1.14
8	-	-	-	-	0	0.36
9	-	-	-	-	0	0.11
10	-	-	-	-	1	0.09
	$a = 1.0576$; $\alpha = 0.0994$; $X_3^2 = 5.858$; $P = 0.1187$.		$a = 0.9941$; $\alpha = 0.4181$; $X_1^2 = 0.091$; $P = 0.7635$.		$a = 0.9159$; $\alpha = 0.3333$; $X_4^2 = 2.684$; $P = 0.6120$.	

Text 7: Die Aufregung vergeht, der Berg der Reiseinfuhren bleibt (Gyôretsu kiete unyû kome no yama). (aus: Asahi Shinbun, Nr. 38861, 1994, S.1. Sparte nicht gekennzeichnet).

Text 8: Gemeinschaftsarbeit in den Filmgesellschaften (...eiga kaisha ni gassaku...). (aus : Asahi Shinbun, Nr. 38691, 1993, S.17, Sparte „Kultur").

Text 9: Eine Freundschaft seit den Vorfahren vor 400 Jahren (Shitashimi-wa 400 nen mae no sosen irai). (aus: Asahi Shinbun, Nr. 38691, 1993, Sparte nicht gekennzeichnet).

	Text 10		Text 11	
x	n_x	NP_x	n_x	NP_x
1	37	37.33	131	130.23
2	29	31.84	85	84.92
3	28	24.06	49	52.30
4	14	12.80	30	22.06
5	6	6.10	5	8.54
6	2	3.87	1	2.78
7	-	-	1	1.17
	$a = 1.1337$; $\alpha = 0.2477$ $X_3^2 = 1.906$; $P = 0.5921$		$a = 0.8410$;$\alpha = 0.2247$ $X_4^2 = 5.691$; $P = 0.2235$	

Text 10: OECD, Mexiko tritt bei (OECD-Mekishiko-ga kamei). (aus: Asahi Shinbun, Nr. 38861, 1994, S.12. Sparte nicht gekennzeichnet).

Text 11: Unnachgiebig: Widerruf des Zusammengehens („Renritsu no kaishô mo to tsuyoki"). (Text aus: Asahi Shinbun, Nr.38861, 1994, S.3. Sparte „Politik/ Gesellschaft").

4. Die Untersuchung hat nun folgendes Ergebnis erbracht: Alle 11 bearbeiteten Pressetexte folgen der Hirata-Poisson-Verteilung. Das Japanische zeigt damit das gleiche Wortlängenverteilungsmodell, das auch für das Französische nachgewiesen werden konnte (vgl. Dieckmann & Judt, 1996; Feldt, Janssen & Kuleisa, 1997). Es ist aber zu berücksichtigen, daß die Datenbasis noch recht gering ist. Ob die gefundene Verteilung sich als ein Modell für ein größeres Textspektrum des Japanischen erweist, können nur weitere Untersuchungen ergeben.

5. Die Wortlängen des Japanischen fanden in der Forschung auch unter anderen Aspekten als hier Beachtung. So hat Fucks (1968:91) insgesamt 11 Sprachen hinsichtlich ihrer mittleren Wortlänge miteinander verglichen und dabei festgestellt, daß das Japanische zu einer Gruppe von Sprachen mit besonders hoher durchschnittlicher Wortlänge gehört, zusammen mit Griechisch, Russisch, Ungarisch, Latein und Türkisch. Die Häufigkeitsverteilung für Wortlängen im Japanischen zeigt, daß die Kurve bei den ein - bis dreisilbigen Wörtern im Vergleich zu den anderen berücksichtigten Sprachen relativ flach verläuft und nur wenig abfällt (Fucks, 1968:80). Diese Beobachtungen passen gut zu typologischen Daten, die Silnitsky (1993:141) präsentiert: Hier zeigt sich das Japanische hinsichtlich des Syntheseindexes SYN = Zahl der Morpheme / Zahl der Wörter als Sprache mit besonders hohem Wert: SYN = 2.71. Nur Arabisch hat unter 31 berücksichtigten Sprachen einen noch höheren Wert: SYN = 3.14. Auch wenn die Wortlängen unterschiedlich gemessen werden - Fucks (1968) bestimmt die Wortlänge nach der Anzahl der Silben, Silnitsky (1964) nach der Anzahl der Morpheme - , so zeichnet sich doch in beiden Untersuchungen übereinstimmend ab, daß das Japanische eine Sprache mit auffallend hoher durchschnittlicher Wortlänge ist. Dies scheint nicht daran zu liegen, daß im Japanischen extrem lange Wörter häufig sind, sondern eher an dem relativ hohen Anteil mittlerer Wortlängen, wie sie sich in den Pressetexten der vorliegenden Untersuchungen gezeigt haben.

Literatur

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax*. Frankfurt: Athenäum.

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In P. Schmidt (Hg.), *Glottometrika 15* (S. 158-165), Trier: WVT.

Feldt, S., Janssen, M., & Kuleisa, S. (1996). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In diesem Band.

Fucks, W. (1968). *Nach allen Regeln der Kunst*. Stuttgart: Deutsche Verlags-Anstalt.

Lewin, B. (1990). *Abriß der japanischen Grammatik*. Wiesbaden: Harrassowitz.

Sanada, H. (1993). Comparison of Effectiveness of Various Basic Vocabularies. In G. Altmann (Hg.), *Glottometrika 14* (S. 122-138), Trier: WVT.

Silnitsky, G. (1993). Typological Indices and Language Classes: A Quantitative Study. In G. Altmann (Hg.), *Glottometrika 14* (S. 139-160), Trier: WVT.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

Glottometrika 16, 1997, 185-194

Zur Modellierung der Wortlängen im Chinesischen

Jinyang Zhu, Karl-Heinz Best

0. Eine der zentralen Aufgaben der Linguistik besteht darin, die latenten Strukturen und Mechanismen aufzuspüren, die die Ausprägung und die Entwicklung der Systeme der Sprachen beeinflussen. Ein wesentlicher Schritt in diese Richtung ist Köhler (1986) gelungen, als er das Konzept der Selbstorganisation und -regulierung der Sprachen entwickelte. Breiter (1994) hat nun untersucht, wie sich die Zusammenhänge zwischen der Länge der Lexeme und ihrer Frequenz und Polysemie bei verschiedenen Wortarten und in verschiedenen Sprachstilen des Chinesischen darstellen und kommt u.a. zu den Feststellungen, „that more frequent hyperlexemes are shorter in general than the less ones" (Breiter, 1994:127) und „the polysemy of shorter hyperlexemes is higher than the polysemy of hyperlexemes of greater length for each lexico-grammatical class." (Breiter, 1994:230) „Hyperlexem" bestimmt sie als „central unit of the lexical system which consists of several lexemes united on the basis of the same set of lexico-semantic components at least in one of the meanings of each lexeme, and the same phonetic (or graphic) representation" (Breiter, 1994:226). Die Ergebnisse Breiters bestätigen für das Chinesische die Richtigkeit wesentlicher Annahmen in Köhlers Regelkreis.

1. Mit einem weiteren Aspekt, dem Zusammenhang zwischen Wortlänge und Häufigkeit, befassen wir uns in der vorliegenden Untersuchung. Den theoretischen Rahmen unserer Arbeit bildet die Theorie der Wortlängenverteilung, wie sie in Wimmer u.a. (1994) entwickelt ist. Dort wird die These vertreten, daß die Wortlängenhäufigkeiten nicht zufällig verteilt seien, sondern bestimmten Gesetzmäßigkeiten folgen. Die Grundidee dabei ist, daß die Häufigkeit, mit der Wörter der Länge x im Text vorkommen, proportional zu der Häufigkeit der Wörter der Länge x - 1 ist:

$$P_2 \sim P_1.$$

Wenn diese Abhängigkeit linear wäre, würde gelten:

$$P_2 = a P_1.$$

Viele Untersuchungen (vgl. Hinweise in Best & Altmann, 1996) haben jedoch gezeigt, daß die Proportionen zwischen den verschiedenen Wortlängenklassen nicht konstant sind, sodaß mit einer Funktion $g(x)$ zu rechnen ist, die den Zusammenhang regelt:

$$P_x = g(x) P_{x-1}.$$

Damit ist das allgemeine Prinzip für die Verteilung von Einheiten, die unterschiedliche Längen annehmen können (z.B. Wörter, Sätze), genannt. Die Funktion $g(x)$ kann nun verschiedene Formen (Wimmer & Altmann, 1996:114) annehmen, je nach dem, welchen Randbedingungen die untersuchte Einheit unterliegt. Als derartige Randbedingungen kommen infrage: die bearbeitete Sprache, die Textsorte, der Autor und evt. weitere. So ergibt sich eine Gruppe von Funktionen, die bisher zur Modellierung der Verteilung von Wortlängen in Texten verwendet werden konnten (Wimmer u.a., 1994:102).

2. Ein erster Versuch zum Chinesischen wurde in Zhu & Best (1997) unternommen. Dabei ging es um 12 moderne Kurzgeschichten mehrerer Autoren; es stellte sich heraus, daß an alle bearbeiteten Texte ein Modell, die positive Cohen-Poisson-Verteilung, angepaßt werden konnte, deren Formel wie folgt lautet:

$$P_x = \begin{cases} \dfrac{(1-\alpha)\mathrm{a}}{e^{a}-1-\alpha\mathrm{a}}, & x=1 \\ \dfrac{a^x}{x!\,(e^{a}-1-\alpha\mathrm{a})}, & x=2,3,4,\ldots \end{cases}$$

$$\mathrm{a}>0;\ 0<\alpha<1.$$

Da 12 Texte einer literarischen Textgattung natürlich nicht als repräsentativ für das Chinesische insgesamt gelten können, werden wir uns hier einer weiteren Textklasse, den Briefen, widmen. Um ein möglichst homogenes Textkorpus zu erhalten, wurden 20 Briefe eines einzigen Verfassers, in diesem Falle Mao Zedongs, ausgewählt. Die zwanzig Briefe stammen aus der Sammlung „Ausgewählte Briefe von Mao Zedong" (Beijing, Volksverlag 1983). Die Texte wurden nach folgenden Kriterien ausgewählt: 1. Die Briefe sollten möglichst neueren Datums sein; 2. sie sollten nicht weniger als 50 Wörter haben; 3. Briefe, die im klassischen Stil geschrieben sind, wurden nicht berücksichtigt.

Die Bearbeitung der Texte erfolgte in der gleichen Weise, wie schon in Zhu & Best (1996) beschrieben. Es wurde für jeden Brief gesondert die Zahl der einsilbigen, zweisilbigen usw. Wörter festgestellt. Das „Wort" wird als eine distributionell-semantische Einheit bestimmt; die Zahl der Silben pro Wort bemißt sich nach der Zahl der in ihm enthaltenen Vokale. Es wird immer nur der laufende Text ausgewertet.

3. In den folgenden Tabellen finden sich die Ergebnisse der Anpassung der positiven Cohen-Poisson-Verteilung an die Daten der Mao-Briefe. Dabei bedeuten: x die Wortlängen, n_x die im jeweiligen Brief beobachtete Häufigkeit, mit der Wörter der Länge x im Brief vorkommen; NP_x die berechnete Häufigkeit dieser Wortlängenklasse. a und α sind die Parameter der Verteilung. X^2 ist das Chiquadrat, FG die Freiheitsgrade; P ist die Überschreitungswahrscheinlichkeit des Chiquadrats, C der Diskrepanzkoeffizient. P wird als zufriedenstellend angesehen, wenn $P \geq 0.05$ ist. In den meisten Fällen kann aber P nicht bestimmt werden, da bei der Anpassung des Modells keine Freiheitsgrade bleiben; in diesen Fällen kann als Prüfkriterium nur C verwendet werden, dessen Wert dann als zufriedenstellend gewertet wird, wenn $C \leq 0.02$ ist. Die Anpassung wurde mit dem Altmann-Fitter (1994) durchgeführt.

Die Ergebnisse:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	44	-	56	56.09	36	35.98
2	38	-	41	40.75	84	84.02
3			3	3.16	4	4.77
4					1	0.23
			$a = 0.2193$ $\alpha = 0.8491$ $X^2 = 0.009$ $C = 0.0001$		$a = 0.1706$ $\alpha = 0.9635$ $X^2 = 0.000$ $C = 0.0000$	

Text 1: Brief 1 vom 20.4.57
Text 2: Brief 2 vom 17.12.57
Text 3: Brief 3 vom 12.1. 58

Anmerkung: an Brief 1 kann kein Modell angepaßt werden, da er nur zwei Wortlängenklassen aufweist. Die senkrechten Linien in den Tabellen zeigen an, daß die entsprechenden Wortlängenklassen zusammengefaßt wurden.

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	59	58.98	53	52.88	61	60.81
2	54	54.01	47	47.46	69	71.79
3	3	7.21	6	6.03	9	9.39
4	5	0.80	1	0.63	4	1.01
	$a = 0.4009$ $\alpha = 0.7811$ $X^2 = 0.000$ $C = 0.0000$		$a = 0.3813$ $\alpha = 0.7876$ $X^2 = 0.023$ $C = 0.0002$		$a = 0.3927$ $\alpha = 0.8337$ $X^2 = 0.761$ $C = 0.0053$	

Text 4: Brief 4 vom 22.3.58
Text 5: Brief 5 vom 22.5.58
Text 6: Brief 6 vom 11.10.58

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	31	31.09	19	19.29	26	26.00
2	54	53.28	31	32.50	27	27.01
3	7	7.63	1	6.20	3	4.39
4			7	0.88	2	0.60
5			0	0.10		
6			1	0.03		
	$a = 0.3882$ $\alpha = 0.8867$ $X^2 = 0.060$ $C = 0.0007$		$a = 0.5725$ $\alpha = 0.8301$ $X^2 = 0.522$ $C = 0.0089$		$a = 0.4875$ $\alpha = 0.7653$ $X^2 = 0.000$ $C = 0.0000$	

Text 7: Brief 7 vom 10.58
Text 8: Brief 8 vom 22.10.58
Text 9: Brief 9 vom 28.7.59

Anmerkung: Brief 7 und 14 sind nicht genau datiert.

	Text 10		Text 11		Text 12	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	20	20.03	28	28.04	32	31.97
2	46	45.96	35	34.95	49	49.02
3	4	4.01	0	0.97	2	3.77
4			1	0.04	2	0.24
	$a = 0.2454$ $\alpha = 0.9465$ $X^2 = 0.0000$ $C = 0.0000$		$a = 0.0840$ $\alpha = 0.9663$ $X^2 = 0.009$ $C = 0.0000$		$a = 0.2312$ $\alpha = 0.9246$ $X^2 = 0.000$ $C = 0.0000$	

Text 10: Brief 10 vom 3.8.59
Text 11: Brief 11 vom 29.12.59
Text 12: Brief 12 vom 10.10.60

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	34	33.97	30	30.91	81	80.34
2	43	43.03	27	27.23	69	70.21
3	3	3.74	5	5.06	10	10.22
4	0	0.24	1	0.80	1	1.11
5	1	0.02			1	0.12
	30					
	$a = 0.2608$ $\alpha = 0.8970$ $X^2 = 0.000$ $C = 0.0000$		$a = 0.5575$ $\alpha = 0.6835$ $X^2 = 0.006$ $C = 0.0001$		$a = 0.4367$ $\alpha = 0.7501$ $X_1^2 = 0.528$ $P = 0.47$	

Text 13: Brief 13 vom 5.12.60
Text 14: Brief 14 von 1960
Text 15: Brief 15 vom 20.1.61

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	45	44.83	78	78.23	30	29.98
2	39	39.73	82	82.65	33	32.99
3	4	4.09	7	10.10	2	4.51
4	1	0.35	5	1.02	3	0.52
	$a = 0.3089$ $\alpha = 0.8257$ $X^2 = 0.088$ $C = 0.010$		$a = 0.3669$ $\alpha = 0.8263$ $X^2 = 0.077$ $C = 0.0004$		$a = 0.4105$ $\alpha = 0.8135$ $X^2 = 0.000$ $C = 0.0000$	

Text 16: Brief 16 vom 24.1.61
Text 17: Brief 17 vom 6.5.61
Text 18: Brief 18 vom 14.5.61

	Text 19		Text 20	
x	n_x	NP_x	n_x	NP_x
1	55	55.42	25	25.32
2	53	54.44	18	19.01
3	4	8.13	3	4.66
4	7	1.01	2	0.85
5			1	0.12
6			1	0.04
	$a = 0.4483$ $\alpha = 0.7718$ $X^2 = 0.422$ $C = 0.0035$		$a = 0.7353$ $\alpha = 0.5105$ $X^2 = 0.375$ $C = 0.0075$	

Text 19: Brief 19 vom 22.6.61
Text 20: Brief 20 vom 9.9.61

4. Als Ergebnis dieser Untersuchung kann damit festgestellt werden, daß an 19 der 20 Briefe die positive Cohen-Poisson-Verteilung angepaßt werden kann. Das Scheitern bei dem ersten untersuchten Brief ist auf eine stilistische Besonderheit dieses Textes zurückzuführen: Er weist nur ein- und zweisilbige Wörter auf.

Damit kann die positive Cohen-Poisson-Verteilung weiterhin als ein gutes Modell für chinesische Texte gelten, das bei der Untersuchung weiterer Textsorten und Autoren berücksichtigt werden sollte. Ein ergänzender Hinweis scheint angebracht: An alle 19 Texte, die hier modelliert werden konnten, ließ sich auch die Hyperpoisson-Verteilung erfolgreich anpassen. Diese Tatsache erscheint deshalb als beachtenswert, weil die Hyperpoisson-Verteilung in vielen, auch typologisch sehr verschiedenen Sprachen immer wieder an Texte angepaßt werden konnte, die zur Alltagssprache gehören oder dieser wenigstens nahestehen. An die 12 literarischen Texte (Zhu & Best, 1996) ließ sich die Hyperpoisson-Verteilung nicht mit Erfolg anpassen. Es bleibt weiteren Untersuchungen vorbehalten, hier ein genaueres Bild des Chinesischen zu gewinnen: Lassen sich alle Texte mit nur einer Verteilung modellieren oder müssen für verschiedene Textsorten, Autoren und Zeitabschnitte auch verschiedene Modelle gefunden werden?

5. Es bietet sich an, im Anschluß an die vorliegenden Textanalysen noch kurz auf stilistische und typologische Aspekte einzugehen. In ihrem eingangs genannten Aufsatz gibt Breiter (1994:230) die durchschnittliche Wortlänge für 4 verschiedene Stile des Chinesischen an:

„style of fiction“	:	2.0367
„journalistic style“	:	1.9658
„scientific style“	:	1.9311
„colloquial style“	:	1.8303

Diese Daten wurden auf der Basis von Lexika gewonnen (Breiter 1994: 226). Zum Vergleich seien hier entsprechende Werte aus unseren eigenen Untersuchungen genannt: Die durchschnittliche Wortlänge in den 20 Briefen Maos beträgt 1.66; sie schwankt zwischen 1.46 und 2.00. In den Kurzgeschichten (Zhu & Best, 1997) beträgt der Durchschnittswert 1.73 bei einer Schwankung zwischen 1.49 und 2.04. Das Patentgesetz (Best & Zhu, 1994) weist eine mittlere Wortlänge von 2.07 auf. Der Unterschied dieser Werte zu denen Breiters erklärt sich wenigstens zum Teil damit, daß es sich hier um Werte aus Texten statt aus dem Lexikon handelt. Die Kurzgeschichten gehören ja zum „style of fiction“; zwischen der Angabe Breiters mit einer durchschnittlichen Wortlänge von 2.0367 und den hier bearbeiteten Texten von 1.73 liegt doch eine erhebliche Differenz. Es ist aber zu berücksichtigen, daß im Text besonders die kurzen Wörter die am häufigsten verwendeten sind. Betrachtet man Briefe einmal als dem „colloquial style“ zugehörig, so ergibt sich ebenfalls eine beträchtliche Differenz zwischen

Breiters Angabe (1.8303) und dem hier gefundenen Durchschnittswert (1.66). Das Patentgesetz ist wohl am ehesten dem „scientific style" zuzuordnen; die Differenz zwischen Breiters Angabe (1.9311) und dem Patentgesetz (2.07) weist ebenfalls eine deutliche Differenz auf, erstaunlicherweise aber in anderer Richtung: Der untersuchte Text zeigt einen höheren Durchschnittswert als den, den Breiter aufgrund des Lexikons erhoben hat. Nun ist natürlich zu beachten, daß einzelne Texte, evt. auch bestimmte Textgruppen, sich deutlich vom Durchschnitt ihrer Textsorte oder ihres Funktionalstils abheben können. Es mag also eine Besonderheit der Gesetzessprache oder auch nur dieses einen Gesetzestextes vorliegen.

Immerhin zeigen Breiters Angaben ebenso wie unsere an einzelnen Texten gewonnenen Durchschnittswerte für die Wortlängen deutlich, daß verschiedene Stile in einer Sprache sich deutlich hinsichtlich eines bestimmten Kriteriums unterscheiden können und daß man sehr verschiedene Werte erhalten kann, wenn man Daten auf der Grundlage von Lexika oder von Texten erhebt. Hinsichtlich des letztgenannten Aspekts haben Grotjahn & Altmann (1993:143-6) betont, daß Texte und Wörterbücher als Datengrundlage sorgfältig zu unterscheiden sind.

Diese Beobachtung eröffnet erstens eine Chance, auf statistischer Grundlage Unterschiede zwischen verschiedenen Stilen innerhalb einer Sprache zu erfassen; sie ermöglicht aber zweitens, ein Grundproblem der quantitativen Typologie, das seit Greenbergs bahnbrechender Arbeit (Greenberg, 1960) besteht, wenigstens teilweise zu lösen. Dieses Problem besteht darin, daß zur typologischen Charakterisierung einer Sprache Indizes auf der Basis eines sehr kurzen Textabschnittes für ausreichend gehalten wurden. Der Blick auf die Angaben zur durchschnittlichen Wortlänge im Chinesischen zeigt jedoch, daß mit erheblichen Schwankungen dieses Wertes zu rechnen ist.

Vergleicht man damit neue Angaben zum Arabischen, so zeigen sich wiederum sehr unterschiedliche Synthesewerte: So gibt Silnitsky (1993:141) für S = M/W den Wert 3.14 an; Stepanov (1995:147) dagegen 2.58 ± 0.03, also einen erheblich niedrigeren Wert. (S: Synthese-Indes; M: Zahl der Morphe im Text; W: Zahl der Wörter im Text.)

Die genannten Synthesewerte des Chinesischen und des Arabischen erweisen sich als derart unterschiedlich, daß es sich in beiden Fällen eigentlich um Sprachen verschiedenen Typs handeln müßte, wenn man Greenbergs Vorschlag folgt, der bei S = 1.00 - 1.99 von analytischen, bei S = 2.00 - 2.99 von synthetischen und bei S > 3.00 von polysynthetischen Sprachen spricht (Greenberg, 1960:194).

Der Blick auf beide Sprachen zeigt damit eindringlich, daß es sich hierbei um ein generelles Problem handelt. Daraus muß man wohl folgende Konsequenzen ziehen:

Bei der Untersuchung von Sprachen zu typologischen Zwecken sollte man darauf achten, daß die Indizes aus Texten vergleichbarer Textsorten gewonnen

werden. Dabei ist es sinnvoll, mehrere Texte der fraglichen Textsorten zu bearbeiten, um nicht zufällig einen extrem ausgeprägten Text zu berücksichtigen. In unseren Kurzgeschichten weist z.B. einer der Texte eine durchschnittliche Silbenlänge der Wörter von 2.04 auf; alle andern Texte zeigen einen weit niedrigeren Wert. Solche „Ausreißer" finden sich bei niedrigen ebenso wie bei hohen Werten und in vielen Textgruppen. Vielleicht sollte man daraus die Konsequenz ziehen, für typologische Zwecke Extremwerte eines Datensatzes auszuschließen. Will man darüber hinaus auch noch eine Vorstellung von der Variabilität einer Sprache gewinnen, wird man wohl nicht nur eine, sondern mehrere, möglichst unterschiedliche Textsorten oder Funktionalstile bearbeiten müssen. Der Aufwand gegenüber dem in Greenberg (1960) vorgestellten Verfahren erhöht sich damit erheblich; man kommt aber so dem Ziel näher, tatsächlich auch Vergleichbares miteinander zu vergleichen und nicht zufällige Extremwerte zur Grundlage der typologischen Arbeit zu machen.

Literatur

Best, K.-H., & Altmann, G. (1996). Project Report. *Journal of Quantitative Linguistics*, 3, 85-88.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II*, (S. 19-30), Stuttgart: Steiner.

Breiter, M.A. (1994). Length of Chinese Words in Relation to their Other Systemic Features. *Journal of Quantitative Linguistics*, 1, 224-231.

Greenberg, J.H. (1960). A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics*, 26, 178-194.

Grotjahn, R., & Altmann, G. (1993). Modelling the Distribution of Word Length: Some Methodological Problems. In R. Köhler & B. Rieger (Hg.), *Contributions to Quantitative Linguistics* (S. 141-153). Dordrecht: Kluwer.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Silnitsky, G. (1993). Typological Indices and Language Classes: A Quantitative Study. In G. Altmann (Hg.), *Glottometrika 14* (S. 139-160), Trier: WVT.

Stepanov, A.V. (1995). Automatic Typological Analysis of Semitic Morphology. *Journal of Quantitative Linguistics*, 2, 141-150.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Zhu, J., & Best, K.-H. (1997). Wortlängenhäufigkeiten in chinesischen Kurzgeschichten. *Asian and African Studies* (im Druck).

Software

Altmann-FITTER. (1994). Lüdenscheid: RAM-Verlag

Glottometrika 16, 1997, 195-203

Wortlängenhäufigkeiten im Ungarischen

Hans-Hermann Bartens, Thomas Zöbelin

0. Gegenstand dieser Untersuchung ist die Frage, mit welcher Häufigkeit Wörter unterschiedlicher Silbenzahl in ungarischen Texten vorkommen und ob sich für diese Häufigkeiten so wie in anderen Sprachen auch Gesetzmäßigkeiten nachweisen lassen. Soweit uns bekannt, wurde dieses Problem im Ungarischen nur von Fónagy (1960, 1961) behandelt. Nachdem bisher hauptsächlich indoeuropäische Sprachen bearbeitet wurden (vgl. z.B. die Beiträge in Glottometrika 15), kommt hier nach dem Türkischen (Altmann, Erat & Hřebíček, 1996) eine weitere agglutinierende Sprache zur Geltung.

Die ungarischen Texte sind aus verschiedenen Textsorten gewählt: Es handelt sich um literarische (Lyrik und Prosa) und pressesprachliche Texte. Das Textkorpus ist damit von vornherein relativ heterogen.

1. Wie in den anderen Arbeiten wird das „Wort" als orthographisches Wort bestimmt; die Zahl der Silben pro Wort bemißt sich nach der Zahl der Vokale im Wort. Bei den nullsilbigen Wörtern handelt es sich stets um die Variante „s" der Konjunktion „és" 'und'. Bei Kurzwörtern wurde die Silbenzahl gemäß API ermittelt; Abkürzungen wurden in ihrer ausgeschriebenen Form, Fremdwörter gemäß ihrer jeweiligen Aussprache analysiert. Auch bei Zahlwörtern wurde deren ausgeschriebene Form berücksichtigt. Fälle wie „Entzugs- und Entgiftungsplätze" im Deutschen wurden in den ungarischen Texten so wie im Deutschen behandelt: Im angegebenen Beispiel als 3 Wörter mit 2/1/5 Silben.

2. An die ermittelten Daten wurden zwei Verteilungen angepaßt:

a) die positive Poisson-Verteilung

$$P_x = \frac{a^x}{x!(e^a - 1)}, \quad x = 1, 2, 3, \ldots$$

b) die positive Singh-Poisson-Verteilung

$$P_x = \begin{cases} 1 - a + \dfrac{\alpha a e^{-\alpha}}{1-e^{-\alpha}}, & x=1 \\ \dfrac{\alpha a e^{-\alpha}}{x!(1-e^{-\alpha})}, & x=2,3,... \end{cases}$$

Die Ergebnisse sind in den folgenden Tabellen zu sehen. Zu den Texten werden folgende Werte angegeben:

x - Wortlänge, gemessen in der Zahl der Silben;
f_x - Zahl der Wörter des jeweiligen Textes mit Länge x;
NP_x - theoretische Werte der positiven Poisson-Verteilung;
NP_{x1} - theoretische Werte der positiven Singh-Poisson-Verteilung;
a, α - Parameter;
X_k^2 - Wert des Chiquadrats mit k Freiheitsgraden;
P - Überschreitungswahrscheinlichkeit des Chiquadrats;
$C = X^2/N$ - Diskrepanzmaß, das vor allem bei großer Gesamtzahl (N) der Wörter benutzt wird.

Das Ergebnis des Anpassungstests wird dann als zufriedenstellend betrachtet, wenn $P \geq 0.05$ oder - besonders bei großem N - $C \leq 0.02$.

Die Kürzung der Konjunktion „és" auf „s" demonstriert, wie Erweiterungen und Modifikationen der empirischen Verteilungen entstehen. In solchen Fällen wäre es natürlich angebrachter, wenn wir die beiden gestutzten Verteilungen auf die Klasse $x = 0$ erweitert hätten. Dies hätte aber die Zugabe eines weiteren Parameters erfordert. Da die Häufigkeiten in Klasse $x = 0$ jedoch immer sehr klein waren, zogen wir es vor, sie mit der Klasse $x = 1$ zusammenzufassen. Ein ähnliches Phänomen kann man in slawischen Sprachen beobachten, wo durch Wegfall von „jers" nullsilbige Präpositionen entstanden sind. In solchen Fällen ist eine Erweiterung des Modells angebracht (vgl. Uhlířová).

	Text 1		Text 2		
x	n_x	NP_x	n_x	NP_x	NP_{x1}
0	-	-	8	-	-
1	82	97.71	68	85.07	75.05
2	109	94.50	90	78.02	88.36
3	73	60.92	49	47.70	52.19
4	21	29.46	26	21.87	20.55
5	12	11.39	1	8.02	6.07
6	2	5.02	2	3.32	1.78
	$a = 1.9342$; $X_3^2 = 9.92$; $P = 0.02$.		$a = 1.8343$; $X_2^2 = 3.37$; $P = 0.19$.		$a = 1.1814$; $\alpha = 0.9990$; $X_1^2 = 5.95$; $P = 0.11$.

Text 1: Balassi, Bálint (1554-1594): Bocsásd meg, Úristen... (Verzeih, Herr...) (anderweitig auch u.d.T.: Kiben bűne bocsánatáért könyörgett) (1584) (Text aus: Magyar versek könyve. Hrsg. János Horváth. Budapest: Magyar Szemle Társaság 1937, S. 46 - 48).
Text 2: Balassi, Bálint: Katonaének (Soldatenlied) (auch: Egy katonaének)(1589) (Textquelle wie Text 1, S. 52f).

	Text 3			Text 4		
x	n_x	NP_x	NP_{x1}	n_x	NP_x	NP_{x1}
0	-	-	-	3	-	-
1	28	29.47	29.83	81	83.85	83.96
2	31	27.94	27.23	65	71.66	71.67
3	16	17.66	17.61	54	40.83	40.79
4	12	8.37	8.54	15	17.45	17.41
5	1	4.56	4.79	1	5.96	5.94
6	-	-	-	3	2.25	2.23
	$a = 1.8962$ $X_2^2 = 0.57$ $P = 0.75$.	$a = 1.9402$ $\alpha = 0.9800$ $X_2^2 = 5.16$ $P = 0.08$.		$a = 1.7094$ $X_3^2 = 7.35$ $P = 0.06$		$a = 1.7074$ $\alpha = 0.99$ $X_2^2 = 7.35$ $P = 0.03$.

Text 3: Balassi, Bálint: Ez világ sem... (Die Welt nicht...) (auch: Hogy Juliára tatála) (1588) (Textquelle wie Text 1, S.54).
Text 4: Balassi, Bálint: A fülemiléhez (An die Nachtigall) (auch: A fülemilének szól) (1588) (Textquelle wie Text 1, S. 54f).

	Text 5			Text 6		
x	n_x	NP_x	NP_{x1}	n_x	NP_x	NP_{x1}
0	-	-	-	3	-	-
1	43	41.77	42.20	121	127.34	127.34
2	27	30.99	30.32	105	103.59	103.59
3	21	15.32	15.33	61	56.17	56.18
4	5	7.92	8.15	25	22.85	22.85
5	-	-	-	4	7.43	7.43
6	-	-	-	1	2.62	2.62
	$a = 1.4837$; $X_2^2 = 3.71$; $P = 0.16$.		$a = 0.5169$; $\alpha = 0.9769$; $X_1^2 = 3.69$; $P = 0.05$.	$a = 1.6270$; $X_4^2 = 3.30$; $P = 0.51$.		$a = 1.6271$; $\alpha = 0.99$; $X_3^2 = 3.30$; $P = 0.35$.

Text 5: Csokonai Vitéz, Mihály (1773-1805): A rózsabimbóhoz (An die Rosenknospe) (1803) (Textquelle wie Text 1, S. 180f).

Text 6: Csokonai Vitéz, Mihály: Szerelemdal a csikóbőrös kulacshoz (Liebeslied an die mit Fohlenfell bezogene Feldflasche) (1802) Textquelle wie Text 1. S. 188f).

	Text 7			Text 8		
x	n_x	NP_x	NP_{x1}	n_x	NP_x	NP_{x1}
0	3	-	-	7	-	-
1	39	43.55	42.41	74	76.66	80.88
2	36	36.22	36.16	64	71.94	66.38
3	22	20.08	20.55	48	45.00	44.06
4	10	8.35	8.76	31	21.11	21.94
5	1	2.77	2.98	2	11.29	12.74
6	1	1.03	1.14	-	-	-
	$a = 1.6634$; $X_3^2 = 1.40$; $P = 0.71$.		$a = 1.7052$; $\alpha = 0.99$; $X_3^2 = 1.62$; $P = 0.66$.	$a = 1.8767$; $X_2^2 = 1.33$; $P = 0.51$.		$a = 1.9915$; $\alpha = 0.9371$; $X_1^2 = 0.52$; $P = 0.47$.

Text 7: Csokonai Vitéz, Mihály: A magánossághoz (An die Einsamkeit) (1798) (Textquelle wie Text 1, S. 189).

Text 8: Petőfi, Sándor (1823 - 1849): Az alföld (Das Tiefland) (1844) (Textquelle wie Text 1, S. 360).

	Text 9			Text 10		
x	n_x	NP_x	NP_{x1}	n_x	NP_x	NP_{x1}
0	6	-	-	-	-	-
1	117	122.18	122.18	57	57.84	58.69
2	86	92.12	92.12	52	50.56	49.44
3	56	46.30	46.30	29	29.47	29.33
4	17	17.45	17.45	19	19.13	19.54
5	3	6.95	6.95	-	-	-
	$a = 1.5079$; $X_3^2 = 4.69$; $P = 0.20$.		$a = 1.5079$; $\alpha = 0.99$; $X_2^2 = 4.69$; $P = 0.10$.	$a = 1.7484$; $X_2^2 = 0.06$; $P = 0.97$.		$a = 1.7799$; $\alpha = 0.9800$; $X_1^2 = 0.20$; $P = 0.66$.

Text 9: Petőfi, Sándor: A puszta, télen (Die Pußta im Winter) (1848) (Textquelle wie Text 1, S. 392f).

Text 10: Petőfi, Sándor: Nemzeti dal (Nationallied) (1848) (Textquelle wie Text 1, S. 395f).

	Text 11		
x	n_x	NP_x	NP_{x1}
0	1	-	-
1	64	61.94	62.95
2	58	55.81	54.53
3	24	33.52	33.33
4	27	22.73	23.19
	$a = 1.8020$; $X_2^2 = 3.75$; $P = 0.15$.		$a = 1.8338$; $\alpha = 0.9800$; $X_1^2 = 3.53$; $P = 0.06$.

Text 11: Petőfi, Sándor: Szülőföldemen (Auf meiner Heimaterde) (1848) Textquelle wie Text 1, S. 396).

Prosa (liter. Texte):

	Text 12		Text 13	
x	n_x	NP_x	n_x	NP_x
0	2	-	39	-
1	244	250.56	696	736.94
2	229	221.60	565	564.52
3	159	130.65	285	288.30
4	40	57.77	130	110.42
5	7	20.43	23	33.83
6	8	7.99	6	8.64
7	-	-	1	2.35
	$a = 1.7688$; $X_1^2 = 0.37$; $P = 0.54$.		$a = 1.5321$; $X_5^2 = 8.55$; $P = 0.13$.	

Text 12: Móra, Ferenc (1879-1934): A didergő király (Der fröstelnde König) (Text aus: Régi szép mesék (Alte schöne Märchen). Budapest: Offset és Játékkártya Nyomda o.J., ohne Pag).

Text 13: Móricz, Zsigmond (1879-1942): Judith és Eszter (Judith und Esther) (Text aus: Móricz, Zsigmond: Válogatott novellái (Ausgewählte Erzählungen). Budapest: Szépirodalmi Könyvkiadó 1988, S. 49-55).

	Text 14			Text 15		
x	n_x	NP_x	NP_{xl}	n_x	NP_x	NP_{xl}
0	2	-	-	3	-	-
1	330	330.88	331.37	287	301.90	301.77
2	265	275.86	275.27	266	264.67	264.65
3	166	153.32	153.25	181	154.68	154.73
4	68	63.91	63.99	65	67.80	67.85
5	21	21.31	21.37	18	23.77	23.80
6	1	7.72	7.75	2	9.18	9.20
	$a = 1.6674$; $X_3^2 = 3.43$; $P = 0.33$.	$a = 1.6702$; $\alpha = 0.9979$; $X_2^2 = 3.43$; $P = 0.18$.		$a = 1.7533$; $X_4^2 = 12.07$; $P = 0.017$; $C = 0.0145$.	$a = 1.7540$; $\alpha = 0.99$; $P = 0.007$; $X_3^2 = 12.07$; $C = 0.0145$.	

Text 14: Csáth, Géza (1888-1919) A tor (Das Mahl) (Text aus: Ismeretlen házban. 1. Novellák, drámák, jelenetek (In unbekanntem Hause. 1. Erzählungen, Dramen, Auftritte). Újvidék: Forum Nyomda 1977, S. 20-23)

Text 15: Csáth, Géza: Fekete csönd (Schwarze Stille) (Textquelle wie Text 14, S. 34-37)

	Text 16			Text 17	
x	n_x	NP_x	NP_{xl}	n_x	NP_{xl}
1	312	298.88	310.64	287	270.93
2	256	286.64	272.75	262	286.33
3	206	183.26	180.42	192	201.74
4	87	87.87	89.51	139	106.60
5	28	33.71	35.52	37	45.06
6	14	10.77	11.75	11	15.87
7	2	3.87	4.41	4	4.79
8	-	-	-	1	1.68
	$a = 1.9180$; $X_5^2 = 9.50$; $P = 0.09$.		$a = 1.9845$; $\alpha = 0.9605$; $X_4^2 = 8.05$; $P = 0.09$.	$a = 2.1137$; $X_6^2 = 16.66$; $P = 0.01$.	

Text 16: Csáth, Géza: Apa és fiú (Vater und Sohn) (Textquelle wie Text 14, S. 99-102).

Text 17: Örkény, István: Családunk szeme fénye (Der Glanz in den Augen unserer Familie) (1955) (Text aus: ders., Egyperces novellák (Einminütige Erzählungen) 3., erw. Aufl. Budapest: Magvető Kiadó 1974, S. 28-33).

	Text 18		Text 19	
x	n_x	NP_x	n_x	NP_x
1	211	198.70	276	274.93
2	142	167.71	218	220.08
3	106	94.37	158	164.98
4	45	39.82	109	92.75
5	14	13.44	38	41.72
6	1	4.96	16	15.63
7	-	-	2	6.91
	$a = 1.6881$; $X_3^2 = 7.43$; $P = 0.06$.		$a = 2.2489$; $\alpha = 0.9030$; $X_4^2 = 6.97$; $P = 0.14$.	

Text 18: Örkény, István: Egy lelkiismeretes olvasó (Ein gewissenhafter Leser) (Textquelle wie Text 17, S. 102-104).

Text 19: Örkény, István: Magyar panteon (Ungarisches Pantheon) (Textquelle wie Text 17, S. 64-68).

Pressetexte:

	Text 20		Text 21	
x	n_x	NP_x	n_x	NP_x
0	15	-	4	-
1	433	413.64	313	316.62
2	362	402.84	248	233.52
3	249	261.55	209	226.05
4	137	127.36	163	164.12
5	60	49.61	93	95.32
6	16	16.10	64	46.14
7	4	4.48	7	19.14
8	0	1.09	5	6.94
9	1	0.33	3	2.24
10	-	-	1	0.65
11	-	-	1	0.26
	$a = 1.9478$; $X_6^2 = 10.66$; $P = 0.10$.		$a = 2.9041$; $\alpha = 0.8598$; $X_3^2 = 2.67$; $P = 0.44$	

Text 20: Megint egy lépéssel távolabb Amerikától (Wieder einen Schritt weiter entfernt von Amerika) (Text aus: Magyar Hírlap, 1.4.1993, S. 24).

Text 21: Környezetvédelmi törvény - A hetedik pecsét (Umweltschutzgesetz - Der siebte Stempel) (Text aus: HVG (Heti Világgazdaság), XV. Jahrgang, Nr.12 vom 20.3.1993, S. 87ff).

	Text 22	
x	n_x	NP_x
0	4	-
1	256	263.33
2	170	170.86
3	183	160.79
4	97	113.48
5	66	64.07
6	30	30.14
7	9	12.15
8	3	4.29
9	1	1.34
10	1	0.38
11	1	0.17
	$a = 2.8231$; $\alpha = 0.8267$; $X_6^2 = 7.49$; $P = 0.28$.	

Text 22: Bekebelezési menetrend (Einverleibungsfahrplan) (Text aus: Heti Világgazdaság), XII. Jahrgang, Nr. 15 (567) vom 6.4.1990, S. 11-13).

3. Es kann nun festgestellt werden, daß alle 22 untersuchten Texte trotz ihrer Heterogenität mit der positiven Poisson-Verteilung modelliert werden können. Im Falle von Text 1, 15 und 17 ist die Anpassung mit $P \geq 0.02$ gerade noch akzeptabel. Die positive Singh-Poisson-Verteilung ist ebenfalls für viele ungarische Texte ein geeignetes Modell, erbringt aber in keinem Fall bessere Ergebnisse als die positive Poisson-Verteilung. Eine Berechnung der Daten für den Text in Veenker (1982:320) bestätigt dieses Ergebnis.

Es ist klar, daß trotz dieses zufriedenstellenden Ergebnisses noch Forschungsbedarf zum Ungarischen besteht: Sowohl weitere Texte der bereits hier behandelten als auch solche ganz anderer Textsorten sollten überprüft werden. Es wird sich dann herausstellen, ob die hier gefundenen Verteilungen für ein noch größeres Textspektrum geeignet sind oder ggfs. eine oder mehrere andere Verteilungen ebenfalls verwendet werden können, wie sich am Beispiel von Text 1 andeutet.

Literatur

Altmann, G., Erat, E., & Hřebíček, L. (1996). Word Length Distribution in Turkish Texts. In P. Schmidt (Hg.), *Glottometrika 15* (S. 195-204), Trier: WVT.

Fónagy, I. (1960). A szavak hossza a magyar beszédben (Die Länge der Wörter in der ungarischen Rede). *Magyar Nyelvőr 1960*, 355-360.

Fónagy, I. (1961). Die Silbenzahl der ungarischen Wörter in der Rede. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschun*, 14, 88-92.

Uhlířová, L. (1996). How long are Words in Czech? In P. Schmidt (Hg.), *Glottometrika 15* (134-146), Trier: WVT.

Veenker, W. (1982). Konfrontierende Darstellung zur phonologischen Statistik der ungarischen und finnischen Schriftsprache. *Nyelvtudományi Közlemények*, 84, 305-348.

Glottometrika 16, 1997, 204-212

Wortlängen in Ketschua-Texten

Karl-Heinz Best, Paulina Medrano

0. Im Rahmen des Projekts zur Erforschung der Häufigkeit, mit der Wörter verschiedener Länge in Texten verwendet werden, konnten bisher Daten zu über 30 Sprachen erhoben und ausgewertet werden (Altmann & Best, 1996). Dabei handelt es sich vor allem um in Europa verwendete indoeuropäische und finnisch-ugrische sowie um einige ostasiatische Sprachen. Mit dieser Arbeit können wir erstmals Daten einer der vielen autochthonen Sprachen des amerikanischen Doppelkontinents vorstellen: des Ketschua. Schon Fr. v. Schlegel (1808) hat auf das Ketschua aufmerksam gemacht und zählt es zu den „mechanischen" Sprachen, den „Sprachen durch Affixa", billigt ihm aber zu, „nicht nur sehr kraft- und ausdrucksvoll, sondern auch verhältnißmäßig gebildet und kunstreich" zu sein. (Schlegel, 1808:304f) Ketschua ist "die wichtigste Sprache Mittel- und Südamerikas" und gehört zum agglutinierenden Typ (Wendt, 1987: 196).

1. Wie bei allen Untersuchungen des Projekts gehen wir von der Annahme aus, daß Wortlängenverteilungen in Texten bestimmten Gesetzmäßigkeiten folgen, wie sie von Wimmer u.a. (1994) theoretisch begründet wurden. Die dort entwickelte Theorie, in Wimmer & Altmann (1996) fortgeführt, sieht vor, daß die Häufigkeit, mit der eine Wortlängenklasse P_x in einem Text vorkommt, proportional zur nächstniedrigeren Wortlängenklasse ist, also:

$$P_x = g(x)\ P_{x-1}.$$

Dabei stellt $g(x)$ seinerseits eine Funktion dar; je nachdem, welche Funktion für $g(x)$ genau angenommen wird, kommt man zu den in Wimmer u.a. (1994) und Wimmer & Altmann (1996) vorgeschlagenen oder auch noch zu weiteren Verteilungsfunktionen. Eine ganze Reihe von Faktoren kann für die Wahl einer bestimmten Verteilungsfunktion eine Rolle spielen; die wichtigsten dürften sein: Sprachtyp, Erscheinungsjahr, Textsorte, Stil und Autor. Um zu einer geeigneten Hypothese zu kommen, welcher der möglichen Verteilungen die Wortlängen im Ketschua entsprechen könnten, haben wir uns von folgenden Überlegungen leiten lassen:

Ketschua ist eine agglutinierende Sprache; in Sprachen dieses Typs spielt die Hyperpoisson-Verteilung offenbar eine besondere Rolle (Bartens & Best, 1996). Es liegt damit die Hypothese nahe, daß dies evt. auch für Ketschua-Texte gelten könnte. Diese Hypothese wird noch dadurch gestützt, daß die Hyperpoisson-Verteilung sich immer wieder - auch bei Sprachen anderer Typen wie z.B. Deutsch - bei literarisch einfachen Formen (etwa: Briefen) bewährt hat. Dazu paßt sehr gut, daß uns zur Untersuchung des Ketschua einerseits Gedichte für Kinder, andererseits Märchen zur Verfügung standen.

2. Es werden insgesamt 13 Gedichte für Kinder und 11 Märchen ausgewertet. Die Bearbeitung erfolgte in der üblichen Weise (Best & Zhu, 1994:20): Es wurden die Wörter verschiedener Länge, die in dem laufenden Text vorkommen, nach ihrer Silbenzahl erfaßt. Als „Wort" wurde das orthographische, d.h. die durch Leerstellen bzw. Interpunktionszeichen isolierte Graphemkette aufgefaßt; die Zahl der Silben im Wort bestimmte sich durch die Zahl der in ihm enthaltenen Vokale und Diphthonge. Entsprechend der o.a. Hypothese wurde an die so gewonnenen Daten mit Hilfe des Altmann-Fitters (1994) die 1-verschobene Hyperpoisson-Verteilung angepaßt, deren Formel so lautet:

$$P_x = \frac{a^{x-1}}{b^{(x-1)}\ {}_1F_1\ (1; b; a)}, \quad x = 1, 2, 3...$$

Dabei sind a, b die Parameter der Verteilung; X^2 ist das Chiquadrat; der Index n zu $X^2{}_n$ gibt die Freiheitsgrade an. X ist die Wortlängenklasse, n_X die beobachtete Häufigkeit der Wörter der Längenklasse x, NP_X die aufgrund der verschobenen Hyperpoisson-Verteilung berechnete Häufigkeit dieser Klasse. Die Anpassung an die einzelnen Texte gilt als gelungen, wenn $P \geq 0.05$; noch akzeptabel sind Anpassungen mit $0.01 \leq P < 0.05$. Wenn P mangels Freiheitsgraden nicht berechnet werden kann, wird der Diskrepanzkoeffizient C zur Beurteilung der Güte der Anpassung verwendet, der die Bedingung $C \leq 0.02$ erfüllen muß.

In einigen Fällen (Texte 1, 2, 4-6) erbringt die Anpassung der 1-verschobenen Hyperpoisson-Verteilung keine akzeptablen Ergebnisse. Der Grund dafür ist, daß - wie Text 5 besonders eindrucksvoll zeigt - mehrere Häufigkeitsgipfel zu beobachten sind. Dem kann man mit einer Modifikation der Verteilung begegnen, indem mittels eines weiteren Parameters a ein Teil der berechneten Häufigkeiten von den 3- auf die 4-silbigen Wörter verlagert wird. Die Formel für diese Modifikation lautet:

$$P_x = \begin{cases} P'_x, & x = 1,2,5,6,... \\ P'_x(1-\alpha), & x = 3, \\ P'_x + P'_{x-1}\alpha, & x = 4, \end{cases}$$

wobei P'_x wiederum die 1-verschobene Hyperpoisson-Verteilung ist.

3. Die Ergebnisse der Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die Ketschua-Texte finden sich in den folgenden Tabellen:

1. Gedichte

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	5	4.44	1	0.84	2	2.26
2	17	12.10	7	5.86	8	10.10
3	11	9.35	4	4.96	14	13.81
4	18	18.83	13	12.03	16	11.15
5	8	7.72	7	6.42	7	6.38
6	6	4.89	4	3.32	1	4.30
7			1	1.39		
8			1	0.70		
	$a = 1.9477$; $b = 0.5728$; $\alpha = 0.5$; $X_2^2 = 0.90$; $P = 0.64$.		$a = 2.2320$; $b = 0.3190$; $\alpha = 0.5$; $X_2^2 = 0.71$; $P = 0.70$.		$a = 1.9705$; $b = 0.4414$; $X_3^2 = 5.153$; $P = 0.16$.	

Text 1: Qhepaman Ch'usaj (Reise in die Vergangenheit). In: Torres 1985, S. 4
Text 2: Llinp'ej wawa (Maler Kind). Textquelle wie 1, S. 16
Text 3: Q'ellunchukunaj wisq'aynin (Käfig des Kanarienvogels). Textquelle wie 1, S. 32

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	1	1.10	3	2.54	6	3.73
2	10	11.05	18	15.24	10	13.62
3	8	7.34	5	5.29	9	8.65
4	20	14.13	20	15.00	25	17.63
5	5	5.09	6	5.00	6	7.34
6	0	1.88	1	1.69	2	3.15
7	0	0.56	0	0.46	0	1.11
8	1	0.18	1	0.13	2	0.44
	$a = 1.5334$; $b = 0.1533$; $\alpha = 0.5$; $X_2^2 = 3.60$; $P = 0.17$.		$a = 1.4339$; $b = 0.2390$; $\alpha = 0.7$; $X_2^2 = 2.50$; $P = 0.29$.		$a = 1.9486$; $b = 0.5341$; $\alpha = 0.5$; $X_3^2 = 6.23$; $P = 0.10$.	

Text 4: Askamitarayku (Der Augenblick). Textquelle wie 1, S. 44
Text 5: Wasillank'aj Lloqallita (Ein Kind, Bauarbeiter der Welt). Textquelle wie 1, S. 56
Text 6: P'isqetujta munakuynin (Romantisches Vögelchen). Textquelle wie 1, S. 68

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	3	2.67	2	9.00	5	14.00
2	24	21.39	30	16.43	39	25.27
3	17	23.71	14	17.71	26	27.66
4	19	14.12	10	13.54	20	21.72
5	7	8.11	9	8.02	4	13.30
6			2	3.88	16	6.67
7			0	1.58	2	2.83
8			4	0.84	1	1.55
	$a = 1.2872$; $b = 0.1609$; $X_2^2 = 4.088$; $P = 0.13$.		$a = 2.6316$; $b = 1.4411$; $X_1^2 = 3.439$; $P = 0.06$.		$a = 2.7798$; $b = 1.5398$; $X_1^2 = 0.879$; $P = 0.35$.	

Text 7: Kantutasta muyuchij (Kantuta Verkäuferin). Textquelle wie 1, S. 48
Text 8: T'akaykusqa K'uku (Süße Frucht des Baumes). Textquelle wie 1, S. 66
Text 9: Qharinchasqa Chajra Runa (Ich bin wie ein Baum auf der Erde). Textquelle wie 1, S. 2

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	2	2.42	5	6.63	3	10.07
2	13	10.43	14	10.65	21	13.75
3	17	15.02	10	10.74	12	12.20
4	12	12.99	6	7.89	3	8.01
5	2	8.03	8	4.56	7	4.17
6	9	6.11	0	2.17	2	1.80
7			0	0.88	0	0.66
8			1	0.48	3	0.34
	$a = 2.1659$; $b = 0.5042$; $X^2_3 = 6.972$; $P = 0.07$.		$a = 2.7113$; $b = 1.6882$; $X^2_4 = 6.799$; $P = 0.15$.		$a = 2.5313$; $b = 1.8542$; $X^2 = 0.005$; $C = 0.000$.	

Text 10: Yakuwan Kutaj (Wasser Mühle). Textquelle wie 1, S. 12
Text 11: Urituqa (Der Papagei). Textquelle wie 1, S. 20
Text 12: Llank'aj mama (Proletarische Mama). Textquelle wie 1, S. 52

	Text 13	
x	n_x	NP_x
1	14	12.09
2	8	12.34
3	12	8.71
4	4	4.70
5	3	3.16
	$a = 2.2908$; $b = 2.2446$; $X^2_2 = 3.177$; $P = 0.20$.	

Text 13: Tatapaj Michij (Die junge Hirtin). Textquelle wie 1, S. 60

2. Märchen

	Text 14		Text 15		Text 16	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	5	4.81	3	12.00	1	1.75
2	51	49.07	44	25.12	22	19.41
3	71	67.76	25	30.63	39	36.79
4	43	50.18	26	26.35	32	38.13
5	24	25.39	12	17.51	31	27.19
6	11	9.75	6	9.48	12	14.78
7	4	3.02	11	4.33	6	6.49
8	1	0.78	0	1.71	5	3.46
9	1	0.24	0	0.59		
10			1	0.28		
	$a = 1.5971$; $b = 0.1566$; $X^2_4 = 2.474$; $P = 0.65$.		$a = 2.9205$; $b = 1.3950$; $X^2_2 = 5.568$; $P = 0.06$.		$a = 2.2866$; $b = 0.2063$; $X^2_5 = 3.599$; $P = 0.61$.	

Text 14: Quwimantawan atuj Kupanmantawan (Der Hase und sein Freund, der Fuchs). In: Nina/Torres, 1992, S. 46
Text 15: Anathuyamanta (... das Wiesel). Textquelle wie 14, S. 47
Text 16: Atujpa jatunp'unchaynin (Der Fuchs und sein Glück). Textquelle wie 14, S. 3

	Text 17		Text 18		Text 19	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	2	2.12	10	10.04	2	13.59
2	30	31.91	34	34.16	43	26.14
3	62	50.34	50	44.31	22	29.31
4	30	41.92	23	35.51	24	23.19
5	26	23.71	25	20.59	9	14.17
6	8	10.15	6	9.35	14	7.06
7	5	3.49	8	3.49	0	2.96
8	1	1.00	1	1.10	2	1.07
9	1	0.36	2	0.45	2	0.51
	$a = 1.7632$; $b = 0.1176$; $X^2_5 = 7.876$; $P = 0.16$.		$a = 2.0973$; $b = 0.6169$; $X^2_3 = 6.571$; $P = 0.09$.		$a = 2.6884$; $b = 1.3978$; $X^2_1 = 2.609$; $P = 0.11$.	

Text 17: Uywajninmantawan Alqunmantawan (Der Herr und der Hund). Textquelle wie 14, S. 9
Text 18: Tatakurakunamantawan wayna sipas sawarakujkunamantawan (Die Pfarrer und die jungen Frauen). Textquelle wie 14, S. 29
Text 19: Chajrayujmantawan Khuchimantawan (Der Bauer und das schlaue Schwein). Textquelle wie 14, S. 23

	Text 20		Text 21		Text 22	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	8	7.30	16	18.94	16	16.10
2	52	47.48	78	64.60	39	39.26
3	60	59.87	78	82.27	44	43.14
4	31	41.79	52	64.41	32	30.61
5	25	20.17	39	36.40	14	16.03
6	7	7.44	18	16.09	6	6.65
7	3	2.22	9	5.84	4	3.21
8	1	0.73	1	2.45		
	$a = 1.5643$; $b = 0.2407$; $X^2_4 = 4.859$; $P = 0.30$.		$a = 2.0320$; $b = 0.5957$; $X^2_5 = 8.798$; $P = 0.12$.		$a = 2.0014$; $b = 0.8211$; $X^2_4 = 0.616$; $P = 0.96$.	

Text 20: Wajchamantawan Qhapajmantawan (Die armen und reichen Menschen) Textquelle wie 14, S. 19

Text 21: Atujmantawan Pumamantawan (Der Fuchs und sein Freund, der Tiger). Textquelle wie 14, S. 15

Text 22: Kawsarimuj nunamanta (Der Teufel und der Tod). Textquelle wie 14, S. 8

	Text 23		Text 24	
x	n_x	NP_x	n_x	NP_x
1	34	32.52	25	25.40
2	84	80.36	77	81.03
3	88	88.53	122	110.46
4	65	62.75	98	95.75
5	22	32.79	45	60.84
6	15	13.57	35	30.51
7	6	4.64	16	12.64
8	1	1.35	3	4.46
9	1	0.34	2	1.91
10	1	0.15		
	$a = 1.9882$; $b = 0.8047$; $X^2_5 = 5.209$; $P = 0.39$.		$a = 2.3806$; $b = 0.7464$; $X^2_6 = 7.630$; $P = 0.27$.	

Text 23: Yuthuwan Atujwan (Der Fuchs und das Rebhuhn). Textquelle wie 14, S. 48

Text 24: Janpirimantawan Quwimantawan (Die Hexe und der Hase). Textquelle wie 14, S. 10

4. Die Ergebnisse der Untersuchung lassen sich wie folgt zusammenfassen:

Die 1-verschobene Hyperpoisson-Verteilung läßt sich an alle 13 Gedichte und 11 Märchen mit Erfolg anpassen, wobei in einigen Fällen eine Modifikation des Modells erforderlich war. Sie kann damit im Sinne der Ausgangshypothese als ein gutes Modell für Ketschua-Texte angenommen werden und unterstützt die Annahme, daß alle sprachlichen Phänomene sich gesetzeskonform verhalten.

Ketschua zeigt einige Besonderheiten der Wortlängenverteilung, wie sie bisher noch kaum beobachtet wurden: Relativ oft hat sich gezeigt, daß einsilbige Wörter nicht die häufigste Wortlängenklasse darstellen. Ein gut dokumentiertes Beispiel dafür ist das Türkische (Altmann, Erat & Hřebíček, 1996; Best & Özmen, 1996); auch in vielen estnischen Texten (Bartens & Best, 1996) zeigt sich dieses Bild.

Ganz ungewöhnlich ist dagegen die Tatsache, daß in etlichen der Ketschua-Texte mehrere Häufigkeitsgipfel auftreten, so daß - optisch betrachtet - der Eindruck einer oszillierenden Verteilung der Häufigkeiten unterschiedlicher Wortlängen entsteht. Dieses Phänomen konnte bisher nur in einigen chinesischen Texten beobachtet werden, die der Alltagssprache relativ fern stehen: Gesetzes-, Presse- und Zeitungstexte (Best & Zhu, 1994:28; Zhu & Best, 1992:52-53). Im Chinesischen kann man hierfür die Tatsache verantwortlich machen, daß beim Aufbau fachsprachlicher Komposita zweisilbige Wörter eine besonders wichtige Rolle spielen (Zhu & Best, 1992:54), ein Aspekt, der als Erklärung der Verhältnisse im Ketschua nicht in Frage kommt, da hier gerade Gedichte besonders auffallen. Es ist beabsichtigt, auf einer breiteren Textgrundlage den Ursachen der Erscheinung auf den Grund zu gehen.

Quellen

Torres, Córdova Mamerto, 1985. *Poemas para ninos*. Sucre: Sucre-Ciudad Universitaria

Nina, Llanos Primitivo/Torres, Córdova Mamerto. 1992. *Qhishwa Jawakuna 2* (Ketschua-Märchen 2). Sucre: Projecto de Revitalización Cultural Quechua TIFAP

Literatur

Altmann, G., Erat, E., & Hřebíček, L. (1996). Word Length Distribution in Turkish Texts. In P. Schmidt (Hg.), *Glottometrika 15* (S. 195-205),Trier: WVT.

Bartens, H.-H., & Best, K.-H. (1996). Wortlängen in estnischen Texten. *Ural-Altaische Jahrbücher N.F. 14* (S. 112-128), (erscheint).

Best, K.-H., & Altmann, G. (1996). Project Report. *Journal of Quantitative Linguistics,* 3,1, 85-88.

Best, K.-H., & Özmen, E. (1996). Wortlängenhäufigkeiten in türkischen Texten und ihre linguistischen Implikationen. *Archiv Orientální,* 64, 19-30.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische. In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Schlegel, Fr. v. (1808). Ueber die Sprache und Weisheit der Indier. In *Friedr. v. Schlegel's sämtliche Werke. Zweite Original-Ausgabe*1848. *Achter Band (S.* 271-319), Wien: Ignaz Klang.

Wendt, H. F. (1987). *Fischer Lexikon Sprachen*. Durchges. u. korrig. Neuausgabe. Frankfurt: Fischer

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics,* 1, 98-106.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: WVT.

Zhu, J., & Best, K.-H. (1992). Zum Wort im modernen Chinesisch. *Oriens Extremus* ,35, 45-60.

Software

Altmann-FITTER. (1994). Lüdenscheid: RAM-Verlag.

Glottometrika 16, 1997, 213-275

Untersuchung zur Satzlängenhäufigkeit im Deutschen

Brigitta Niehaus

1. Einführung

Die Satzlänge gehört zu jenen Merkmalen eines Textes, die vielfach als völlig zufällig aufgefaßt werden und somit als wenig aussagekräftig erscheinen. Dennoch stellen seit etwa 100 Jahren immer wieder Forscher Beobachtungen zu diesem Phänomen an, und es hat sich gezeigt, daß verschiedenartige Problemstellungen existieren, in denen die Kenntnis der Satzlänge eine wichtige Rolle spielt. Von den vielen Fragestellungen, die im Zusammenhang mit der Satzlängenforschung genannt werden können, führen wir im folgenden einige an:

(1) Ist die Satzlänge charakteristisch für den Stil eines Autors oder für ein Genre?
(2) Kann die Satzlänge als Kriterium angesehen werden, strittige oder unbekannte Autorschaft zu klären?
(3) Gibt es eine Veränderung in der Satzlänge eines Sprechers im Laufe seines Lebens?
(4) Gibt es in der Geschichte der Satzlänge von primitiven Formen bis zu komplexen wissenschaftlichen Texten eine Entwicklung?
(5) Welche Faktoren wirken bei der Gestaltung der Satzlänge?
(6) Welche mentalen Prozesse sind während der Satzgenerierung im Gange?
(7) Wie hängt die Satzlänge mit anderen Eigenschaften des Satzes bzw. mit denen der anderen Spracheinheiten zusammen, oder mit anderen Worten, in welchen Regelkreisen spielt Satzlänge eine Rolle?
(8) Inwieweit ist die Satzlänge ein Faktor für den Schwierigkeitsgrad eines Textes?
(9) Existieren mathematische Modelle, mit denen sich die Satzlängenverteilung adäquat beschreiben läßt?
(10) Sind Folgen von Satzlängen chaotische, stochastische oder deterministische Sequenzen?

Obwohl sich die vorliegende Arbeit nur mit dem Aspekt der Modellbildung auseinandersetzt, sollen zumindest einige Problemstellungen kurz erläutert werden, ehe dann eine etwas ausführlichere Darstellung der bereits existierenden Modelle zur Satzlänge und eine Abwägung ihrer Vor- und Nachteile durch die Literatur erfolgt.

Seit Sherman (1888), der als Begründer der Satzlängenforschung gilt (vgl. Altmann, 1988:148), haben immer wieder Forscher die Frage der Satzlänge als Stilcharakteristikum aufgegriffen. Eine wichtige Untersuchung zu diesem Themenkomplex führte Yule (1939) an den Werken dreier englischer Autoren durch, indem er in fortlaufenden Textpassagen die Satzlänge durch die Anzahl der Wörter bestimmte. Die Bestätigung der Vermutung der gleichbleibenden Satzlängen bei Werken eines Autors ergab sich durch den Vergleich einfacher statistischer Werte, wie der durchschnittlichen Satzlänge, dem Median und der Standardabweichung des Yuleschen Koeffizienten oder der Wiederholungsrate[1]. Neben den geringen Schwankungen dieser Werte bei einem Autor lieferte die vergleichende Betrachtung unterschiedlicher Autoren Abweichungen zwischen den Werten, so daß Yule schlußfolgerte, daß die Satzlänge als charakteristisches Merkmal eines Autors angesehen werden kann.

Wake (1957) und Clayman (1981) bestätigten das von Yule gewonnene Ergebnis für griechische Prosa bzw. griechische Hexametergedichte. Anhand von Claymans Arbeit läßt sich zusätzlich aufzeigen, daß die Wahl der Maßeinheit die Resultate beeinflußt, denn sie bestimmte die Satzlänge neben der Anzahl der Wörter zusätzlich auch durch die Anzahl der Silben und Phoneme. Während bei der Messung in Wörtern vergleichbare Ergebnisse zu früheren Arbeiten erzielt wurden, ergaben sich bei den Längenmessungen in Silben oder Phonemen Verteilungen mit mehreren Modi, die eine Beziehung zwischen Satzende und Versende signalisierten.

Im Zusammenhang der Satzlänge als Stilcharakteristikum haben einige Forscher das weiterführende Problem der strittigen oder unbekannten Autorschaft aufgegriffen. Yule (1939) beispielsweise versuchte mit Hilfe des Ergebnisses zur Frage des Stilcharakteristikums die Autorschaft zweier strittiger Werke zu klären, was jedoch nur in einem Fall zu einem relativ sicheren Ergebnis führte.

Auch Mosteller & Wallace (1963) mußten bei ihrer Untersuchung der *Federalist Papers* feststellen, daß sich die von Yule vorgeschlagenen statistischen Merkmale nicht zwingend bei verschiedenen Autoren unterscheiden, denn Madison und Hamilton wiesen nahezu identische durchschnittliche Satzlängen und Standardabweichungen auf. Mosteller & Wallace griffen deshalb auf andere Textmerkmale zur Bestimmung des Autors zurück.

[1] Vgl. G. Herdan (1966): The advanced theory of language in choice and chance. Berlin: Springer.

Morton (1965) und Morton & Mc Leman (1966) beschäftigten sich mit Autorenschaftsfragen von griechischer Prosa. Allgemein stellten sie fest, daß die Satzlängenverteilung in Verbindung mit einfachen statistischen Methoden zur Überprüfung verwendet werden kann, solange keine Dialoge untersucht werden, die Stichproben nicht zu klein sind und die zu vergleichenden Texte zeitlich nicht zu weit auseinander liegen. Indirekt bestätigten sie den Befund von Mosteller & Wallace, indem sie einschränkend darauf hinwiesen, daß mit Hilfe dieser Methode nicht geprüft werden kann, daß zwei Texte vom gleichen Autor stammen, sondern sich nur feststellen läßt, daß sie nicht von demselben Autor verfaßt wurden.

Diese kleine Auswahl an Untersuchungen zeigt bereits, daß die Satzlänge kein eindeutiges Kriterium zur Bestimmung der Autorschaft eines Textes liefert. Es müßten eine ganze Reihe von Texten verschiedener Autoren ausgewertet werden, um ein einigermaßen brauchbares Ergebnis zu erhalten.

Neben den beiden bisher genannten Aspekten gibt es auch Versuche, mit Hilfe der Merkmale eines Textes, die zur Schnelligkeit des Lesens und Verstehens beitragen, Aussagen über den Schwierigkeitsgrad eines Textes zu machen. Die Anforderungen, die ein Text an den Leser stellt, werden dabei zumeist in Form multipler Regressionen, in denen die Satzlänge als ein Regressor betrachtet wird, quantifizierbar gemacht.

Die bekannteste Lesbarkeitsformel ist der *reading-ease*, den Flesch (1948) für englische Texte entwickelte. Er setzte die Anzahl der Wörter pro Satz (*sl*) als Index für die syntaktische Komplexität und die Anzahl der Silben pro 100 Wörter (*wl*) als Index für die Schwierigkeit des Vokabulars an und stellte daraus eine Formel auf, die auf der Skala von 0 bis 100 das Maß der Verständlichkeit angibt.

Eine Vereinfachung der Formel von Flesch stellt die Farr-Jenkins-Peterson-Formel (1951) dar, die statt der Anzahl der Silben nur die Anzahl der einsilbigen Wörter pro 100 Wörter (*nosw*) betrachtet.

Auch für das Deutsche ist man bestrebt, Lesbarkeitsformeln zur Überprüfung der Textschwierigkeit zu verwenden. Eine Übertragung der genannten Formeln auf deutsche Texte ist aber aufgrund des hohen Anteils einsilbiger Wörter im Englischen nicht ohne Modifikation möglich.

In seiner Untersuchung von Hauptschulbüchern der 9. Klasse behielt Mihm (1973) zwar die Flesch-Formel bei, änderte aber die Bewertungsskala, um so auch für das Deutsche gute Bewertungsergebnisse zu erhalten.[2] Sein Aufsatz deckt gleichzeitig die Grenzen der Lesbarkeitsformeln auf, da sie nicht die Schwierigkeit literarischer Texte erfassen können.

Andere Möglichkeiten zur Bestimmung des Schwierigkeitsgrades geben Bamberger & Vanecek (1984), indem die mit Hilfe einer Vielzahl von Faktoren wie der Satz- und Wortlänge, dem Prozentsatz einsilbiger, mehrsilbiger und sel-

[2] Seine Werte lagen zwischen 20 (sehr schwere Texte) und 80 (sehr leichte Texte).

tener Wörter verschiedene Formeln für die Lesbarkeit literarischer Jugendbücher und Sachbücher erarbeiteten. Die relativ große Anzahl der Formeln ergab sich dabei aus dem Versuch, für bestimmte Texte oder Schwierigkeitsstufen bessere Formeln zu entwickeln.[3]

Während bei den bisher genannten Forschern die Beschäftigung mit der Satzlängenverteilung dem Zweck diente, andere Fragen zu klären und praktische Probleme zu lösen, existieren auch Arbeiten, in denen die Autoren die Satzlänge selbst zum Gegenstand ihrer Überlegungen machen und der Frage nachgehen, ob es theoretische Verteilungen gibt, mittels derer sich die Satzlänge modellieren läßt. Zwar gab Yule weder in seinem Aufsatz (1939) noch in seinem Buch (1944) ein mathematisches Modell für die Verteilung der Satzlänge an, doch wies er darauf hin: „They are not of the Poisson type but of the type in which the square of the standard derivation largely exceeds the mean" (Yule, 1939:371).

C.B. Williams (1969) betrachtete statt der Anzahl der Wörter pro Satz die logarithmische Anzahl und erhielt Verteilungen, die denen der Normalverteilung ähnlich waren. Er schlug deshalb die Lognormalverteilung als Modell der Satzlänge vor, wenn die Zufallsvariable durch die Anzahl der Wörter bestimmt ist. Buch (1969) kritisierte den Ansatz Williams, da dieser lediglich ähnliche Texte verglich und keine Testverfahren zur Überprüfung der Hypothese angewendet worden waren. Auch Sichel (1974) war mit Williams vorgeschlagener Verteilung nicht einverstanden. Er verwarf das Modell vor allem auch deshalb, weil es sich bei der Verteilung der Satzlänge um eine diskrete Verteilung handelt, die Lognormalverteilung aber nicht diskret ist. Dieser Einwand erweist sich jedoch aus zwei Gründen als nicht schlüssig, denn zum einen ist die Modellierung diskreter Ereignisse mit stetigen Modellen oder umgekehrt in allen Wissenschaften üblich und zum anderen läßt sich jede stetige Verteilung diskretisieren.

Sichel schlug eine diskrete Verteilung vor, die man erhält, indem man den Parameter λ der Poissonverteilung als eine Zufallsvariable mit der Verteilung

$$f(l) = \frac{1}{2}\frac{\left(2\sqrt{\frac{1-q}{a}}q\right)^{g}}{K_{g}\left(a\sqrt{1-q}\right)}\, l^{g-1}\exp\left[-\left(\frac{1}{q}-1\right)l - a^{2}\frac{q}{4l}\right],$$

betrachtet, wobei $K_\gamma(.)$ die modifizierte Besselfunktion zweiter Art der Ordnung γ ist und $-\infty < \gamma < \infty$, $0 < \theta < 1$ und $\alpha > 0$ gilt. Es ergibt sich damit folgende Verteilung:

[3] Zur Angabe einiger dieser Formeln vgl. R. Bamberger & E. Vanecek, 1984: 82 ff. Zu neueren Arbeiten in dieser Richtung s. J. Tuldava (1993): Measuring text difficulty. Glottometrika 14, 1993, S. 69-81.

$$\phi(r) = \frac{\left(\sqrt{1-\theta}\right)^{\gamma}}{K_{\gamma}(\alpha\sqrt{1-\theta})}\frac{\left(\frac{\alpha\theta}{2}\right)^{r}}{r!}K_{r+\gamma}(\alpha),$$

mit $r = 1,2,...$

Als besonders günstig für die Modellierung der Satzlänge sah Sichel die Verteilung mit $\gamma = -\frac{1}{2}$ an. Die Anwendung dieses Spezialfalls auf Texte aus dem Englischen, Griechischen und Lateinischen zeigte gute Erfolge. Mit Ausnahme eines Textes, dessen Autorschaft umstritten ist, folgten alle Texte dem Modell.

Altmann (1988) war der erste, der in seine Satzlängenforschung die Überlegung einschloß, daß sich die Satzlänge nicht nur durch die Anzahl der Wörter oder noch kleineren Einheiten messen läßt, sondern auch die unmittelbare Komponente des Satzes, d.h. durch den Teilsatz, so operationalisiert werden kann, daß durch sie die Satzlänge bestimmt werden kann. Als Modell der Satz-Clause-Verteilung begründete Altmann die negative Binomialverteilung, für die Verteilung der Satzlänge , bestimmt durch die Anzahl der Wörter, die Hyperpascalverteilung. Diese beiden Modelle bezeichnete er zu Ehren Shermans als *Sherman-Gesetze*.

Altmanns Modelle der negativen Binomialverteilung bzw. der Hyperpascalverteilung lassen sich leicht durch den synergetischen Ansatz herleiten. Altmann betrachtete in seinem Ansatz den Quotienten zweier benachbarter Wahrscheinlichkeiten und berücksichtigte dabei Einwirkungen auf den Text durch den Autor, den Hörer, durch Text- und Ebenenfaktoren. Er zeigt auf, daß das Überspringen der Clause-Ebene und der Rückgriff auf die Wort-Ebene den zusätzlichen Parameter der Textebene nötig machen. Die Ansätze $P_x = \frac{a+bx}{cx} P_{x-1}$ bzw. $P_x = \frac{a+bx}{cx+d} P_{x-1}$ führen dann zur negativen Binomialverteilung bzw. zur Hyperpascalverteilung.

Aufgrund der Berücksichtigung verschiedener Faktoren bei der Herleitung der Modelle erschien Altmann auch Williams Modell ungeeignet, denn es muß von der völligen Zufälligkeit der Satzlängenverteilung ausgegangen werden, will man die Lognormalverteilung als Modell akzeptieren. In allen Wissenschaften betrachtet man zufällige Abweichungen als dem Gaußschen Gesetz folgend und die „Lognormalität" ist nur eine logarithmische Transformation der „Fehler", die „normalverteilt" sein sollten. Jedoch gibt es keinen Nachweis dafür, daß irgendwelche Abweichungen in der Sprache normalverteilt wären. Im Gegenteil: alle Abweichung von einem Modell — das als Gesetz gilt — werden vom Sprecher hervorgebracht und sind daher einseitig, d.h. nicht symmetrisch. In den meisten Fällen gilt für sie das Gaußsche Gesetz nicht einmal asymptotisch.

Die Hyperpascalverteilung als Modell der Satz-Wort-Verteilung wurde anhand von 245 Texten verschiedener Sprachen bestätigt, von denen nur 22 auf der

Ebene $\alpha = 0.05$ signifikant abwichen. Die Satz-Clause-Variante untersuchte Altmann an zehn Texten sieben verschiedener Sprachen, die alle dem Modell der negativen Binomialverteilung folgten.

Die vorliegende Arbeit setzt bei der Untersuchung Altmanns an, da seine Modellierungen die Möglichkeit der linguistischen Interpretation bieten. Untersucht wird die Satz-Clause-Verteilung, deren Modell bislang nur anhand sehr weniger Texte überprüft worden ist. Hilfreich bei der Festlegung der Kriterien zur Datenerhebung ist vor allem die Arbeit von Heups (1980), die sich mit dem Verhältnis von Satzlänge zu Clauselänge beschäftigte und zu dem Resultat kam, daß das Verhältnis der beiden Komponenten durch das Menzerathsche Gesetz gesteuert ist. Es wurde dort die These bestätigt, daß die Clauselänge, gemessen durch die Anzahl der Wörter, kleiner wird, je größer die Satzlänge, gemessen in der Anzahl der Clauses, ist.

2 Die Untersuchung zur Satzlängenverteilung

Zur Überprüfung der Hypothese benötigt man eine umfangreiche Datenbasis. Das folgende Kapitel widmet sich der Beschreibung der Datenerhebung. Es werden dazu Aussagen über die Materialgrundlage gemacht und die Einheiten, mit denen operiert werden soll, eingeführt. Zusätzlich sind einige Regelungen angegeben, wie bei besonderen Konstruktionen vorzugehen ist. Abschließend erfolgt eine Erläuterung des Auszählungsmodus.

2.1 Die Materialgrundlage

Ehe man eine Untersuchung anhand von Texten durchführen kann, ist zunächst festzusetzen, auf welcher Materialgrundlage die Datenerhebung erfolgen soll. Da die Hypothese bislang nur an sehr wenigen Texten überprüft worden ist (vgl. Altmann, 1988:159), erscheint es sinnvoll, zunächst nur einen verhältnismäßig kleinen Textbereich auszuwählen. Die Materialgrundlage beschränkt sich auf Texte unterschiedlicher Funktionalstile mit einer Reiher gemeinsamer Merkmale, die im folgenden festgelegt werden.

Die Wahl einer verhältnismäßig homogenen Textbasis hat den Vorteil, daß die Ergebnisse untereinander vergleichbar sind und sich bei Abweichungen vom Modell eher Gründe nennen lassen als bei der Untersuchung extrem verschiedener Texte.

2.1.1 Die Textbeschaffenheit

Bereits die etymologische Herleitung des Wortes *Text* aus dem lateinischen Verb *textere = weben* (vgl. Sowinski, 1988:31) deutet an, daß Texte nicht eine bloße Aneinanderreihung von Sätzen sind, sondern eine Ganzheit darstellen, bei der

sich alle Teile aufeinander beziehen. Um dieses Beziehungsgeflecht nicht zu untergraben, muß bei den zu untersuchenden Texten stets gewährleistet sein, daß es sich nicht um Auszüge, sondern um abgeschlossene Texte handelt.

Neben der Abgeschlossenheit ist vor allem die Länge der Texte von außerordentlicher Bedeutung. Nach Hammerl „gelten entsprechende statistische Gesetzmäßigkeiten [...] nicht für Texte beliebiger Länge, denn lange Texte sind thematisch, stilistisch und somit auch hinsichtlich sprachlicher Phänomene wesentlich weniger homogen als kürzere Texte oder Textfragmente,"(Hammerl, 1990:155). Untersucht werden deshalb nur Texte mittlerer Länge, die etwa 200 Sätze umfassen. Durch das Längenkriterium, dessen Obergrenze bei 270 Sätzen liegt, sollen Inhomogenitäten aufgrund von Pausen und Unterbrechungen beim Schreiben soweit wie möglich vermieden werden. Als Idealfall für die mathematische Analyse werden Texte angesehen, die „in einem Zug" erzeugt und nicht überarbeitet wurden. Denkbar ist, daß neben sehr langen auch extrem kurze Texte eine andere Verteilung aufweisen, weshalb eine Untergrenze von 140 Sätzen festgelegt wird. Diese Grenzen sind mangels entsprechender Erfahrungen vorläufig nur intuitiv gesetzt. Der zeitliche Rahmen, dem die Texte entstammen müssen, umfaßt die Gegenwart, definiert als Spanne zwischen 1945 und 1994.

Um Homogenitätsproblemen zu entgehen, die bei der Übersetzung fremdsprachlicher Texte ins Deutsche auftreten können, werden nur solche Texte betrachtet, die original deutschsprachig konzipiert worden sind.

2.1.2 Zum Funktionalstil

In den Dreißiger Jahren entstand in der Prager Schule die Theorie der funktionalen Stile, in der davon ausgegangen wird, daß sich Sprache nach ihren Verwendungsbereichen und Mitteilungsfunktionen in verschiedene Teilbereiche gliedern läßt. Im Deutschen unterscheidet man bis zu fünf Funktionalstile:

(1) Sprache des Alltagsverkehrs,
(2) Stil des öffentlichen Verkehrs,
(3) Stil der schönen Literatur,
(4) Stil der Presse und Publizistik,
(5) Stil der Wissenschaft.

Da sich die Sprache des Alltagsverkehrs vornehmlich auf gesprochene Sprache bezieht, in die Untersuchung aber nur schriftlich fixierte Texte aufgenommen werden, bleibt diese Funktionalstil unberücksichtigt. Privatbriefe, die zwar dem Idealfall eines Textes sehr nahe kommen, da sie spontan erzeugt und ohne Pausen niedergeschrieben werden, können ebenfalls nicht untersucht werden, weil sie meist die erforderliche Länge von 200 Sätzen nicht aufweisen. Auch Behörden- und Gesetzestexte, sowie alle weiteren Texte, die dem Stil des öffentlichen

Verkehrs zuzurechnen sind, werden wegen des Längenkriteriums nicht ausgewertet. Die Untersuchung beschränkt sich auf die verbleibenden drei Funktionalstile und untergliedert diese in fünf Textgruppen[4], die aus rein pragmatischen Überlegungen folgendermaßen festgelegt werden:

Dem Stil der schönen Literatur werden zwei Textgruppen zugerechnet: Kurzprosatexte für Kinder und für Erwachsene. Zur Kurzprosa lassen sich alle Texte rechnen, die unter Begriffen wie *Kurzgeschichten, Geschichten, Novellen, Kurzerzählungen* erschienen sind. Im Bereich der Kurzprosa für Kinder werden außerdem nur Texte für die Altersgruppe der Sechs- bis Zwölfjährigen betrachtet, wobei für dieses Kriterium die Altersangaben der Stadtbücherei beachtet werden. Für den Bereich der Presse und Publizistik werden nur Artikel aus dem Nachrichtenmagazin DER SPIEGEL ausgezählt. Berücksichtigt werden können mit Ausnahme von Interviews alle Texte, die die erforderlichen Kriterien erfüllen.

Der Stil der Wissenschaft wird in die Bereiche Fachwissenschaft und Philosophie aufgesplittet, da die Philosophie im Gegensatz zu den Fachwissenschaften keinen eigentümlichen Gegenstand behandelt, sondern grundsätzlich alle Dinge zum Gegenstand philosophischer Reflexion werden können. Als Kriterium für Fachwissenschaft gilt die Publikation der Aufsätze in wissenschaftlichen Zeitschriften. Ausgewertet werden rechtswissenschaftliche, wirtschaftswissenschaftliche und geschichtswissenschaftliche Aufsätze. Texte, die von einem Philosophen geschrieben worden sind, werden als philosophisch angesehen: als Philosoph gilt dabei derjenige, der im Metzler-Philosophen-Lexikon aufgeführt ist (vgl. Lutz, 1989).

Die Untersuchung erfolgt an 85 zufällig ausgewählten Texten mit den genannten Merkmalen, von denen sich jeweils 17 eindeutig einer der Textgruppen zuordnen lassen.. Die Auswahl von 17 Texten pro Textgruppe erscheint günstig zur Überprüfung einer theoretischen Verteilung, denn häufig existieren mehrere Funktionen, die sich als Modell eignen könnten. Bei weniger als 10 Texten pro Textgruppe könnte zumeist nicht entschieden werden, welches Modell das beste ist.

2.2 Definitorische Abgrenzungen

Um die Satzlänge meßbar zu machen, müssen zunächst die sprachlichen Einheiten, mit denen gearbeitet wird, definiert werden. Vor allem geht es um die Bestimmung der Begriffe *Satz, Clause* und *Wort*. Anhand der Vielzahl unterschiedlicher Definitionen - allein zum Satz existieren mehr als 200 (vgl. Bünting & Bergenholtz, 1989:20) - zeigt sich, daß es keine verbindlichen Begriffsbestimmungen gibt. Je nachdem, welche Aspekte thematisiert und in den Vordergrund gestellt sind, ergeben sich unterschiedliche Definitionen. Vor allem die Begriffe

[4] Der Begriff Textgruppe ist in Anlehnung an U. Pieper, 1979:45 gewählt.

Satz und *Wort* werden „in verschiedenen Zusammenhängen theoretischer wie praktischer Art benötigt und entsprechend definiert" (vgl. Bünting & Bergenholtz, 1989:20). Bei einer Satzlängenuntersuchung empfehlen sich Operationalisierungen, die eine Orientierung an äußeren und formalen Merkmalen vornehmen und inhaltliche Aspekte weitgehend unberücksichtigt lassen.

2.2.1 Der Satz

Für die Meßbarkeit der Satzlänge ist weniger die Kenntnis dessen wichtig, was ein Satz ist, als vielmehr die Festlegung von Kriterien, durch die sich Sätze eindeutig gegeneinander abgrenzen lassen. Der Terminus *Satz* wird hierbei nicht als grammatischer Begriff verstanden, sondern bezeichnet immer den „Teil eines konkreten Textes" (vgl. Bünting & Bergenholtz, 1989:27).

Während in der gesprochenen Sprache Satzgrenzen nur sehr schwer oder gar nicht bestimmbar sind, stellt die Abgrenzbarkeit bei schriftlich fixierten Texten keine Schwierigkeit dar. Man gelangt zu folgender Definition des Satzes: „Sätze in Texten sind für uns Einheiten, die durch Satzzeichen eingegrenzt sind." (vgl. Bünting & Bergenholtz, 1989: 27).

Abweichend von Bünting & Bergenholtz, die auch das Semikolon zu den satzabschließenden Zeichen rechnen, wird hier die Annahme vertreten, daß Semikola nicht Sätze, sondern nur komplexe Teilsätze trennen. Als satzabschließend gelten die folgenden Interpunktionszeichen: der Punkt (.), das Fragezeichen (?), das Ausrufungszeichen (!). Der Doppelpunkt (:) nimmt eine Sonderstellung ein, denn er wird nur dann als satzschließendes Zeichen gewertet, wenn das erste Graphem des folgenden Wortes groß geschrieben wird (vgl. Pieper, 1979:46).

Beispiele.:

Sollte es Waigel gelingen, die Fluchtgelder aus der EG zu treiben, gibt es nur einen Gewinner: Die Banken werden prächtig verdienen. = 2 Sätze.
Was richtig ist, ist nur dies: daß unsere Ideen Mächte sind, die unsere Geschichte beeinflussen. = 1 Satz

2.2.2 Der Clause

Ein Clause oder auch Teilsatz[5] bezeichnet eine syntaktische Konstruktion, die eine finite Verbform enthält. Als finit werden diejenigen Verbformen angesehen, die im Gegensatz zum Infinitiv und zum Partizip alle vier grammatischen Kategorien, nämlich Person, Numerus, Modus und Tempus ausdrückt.

[5] Die Begriffe werden in der Arbeit synonym verwendet.

2.2.3 Das Wort

Der Wortbegriff spielt innerhalb der Arbeit keine große Rolle. Da der Terminus jedoch an einigen Stellen verwendet wird, soll er an dieser Stelle der Vollständigkeit halber definiert werden. Die gewählte Wortdefinition berücksichtigt den graphematischen Aspekt. Demnach sind Wörter durch Spatien und Interpunktionszeichen voneinander getrennt, wodurch sich eindeutige Identifizierungen ergeben (vgl. Bünting & Bergenholtz, 1989:39).

2.2.4 Die Satzlänge

Der Einheit, in der die Satzlänge gemessen werden soll, kommt eine entscheidende Bedeutung zu, Unterschiedliche Meßeinheiten führen zu unterschiedlichen Ergebnissen und damit auch zu verschiedenen Gesetzen. Wie in Kapitel 3 aufgezeigt wird, haben bislang die meisten Forscher die Satzlänge durch die Anzahl der Wörter charakterisiert. Sofern sich aber die Ebene unterhalb der zu untersuchenden Größe brauchbar operationalisieren läßt, empfiehlt es sich immer, zunächst das Konstrukt mit Hilfe seiner direkten Konstituenten zu beschreiben, um so die Störfaktoren möglichst gering zu halten. Deshalb soll hier, dem Ansatz Altmanns entsprechend, die Satzlänge durch die Anzahl der Clauses im Satz gemessen werden. Die Länge eines Satzes ist dann praktisch durch die Anzahl der vorhandenen finiten Verbformen bestimmt.

2.3 Besondere Regelungen

Im Verlauf der Datenerhebung traten an den unterschiedlichsten Stellen Probleme auf. Es galt für die Behandlung dieser Zweifelsfälle umfassende und eindeutige Kriterien zu erstellen, um innerhalb des Textkorpus ein einheitliches Auszählungsergebnis zu gewährleisten. Die Festlegungen sind nicht zwingend, solange bestimmte Phänomene einheitlich ausgewertet werden, können auch andere Festsetzungen getroffen werden. Es ergeben sich dann nur systematische Verschiebungen, die keinen Einfluß auf das Ergebnis haben. Wichtig ist nur, daß die kritischen Fälle so geregelt werden, daß sie nicht nach unterschiedlichen Regelungen verschieden bewertet werden können. Bei den einzelnen Entscheidungen wird von den Oberflächenstrukturen ausgegangen, die zugrunde liegenden Tiefenstrukturen bleiben weitgehend unberücksichtigt, um stilistische Eigentümlichkeiten einzelner Autoren nicht zu untergraben.

Im folgenden findet eine Auflistung der wichtigsten Problemfälle statt, die anhand von Textbeispielen illustriert und erläutert werden. Andere, nur vereinzelt auftretende Phänomene wurden ad hoc entschieden und werden hier nicht besprochen. Als Grundregel für diese Entscheidungen galt stets, die Texte möglichst unverändert beizubehalten und somit den Vorgaben des Autors zu folgen.

2.3.1 Direkte und indirekte Rede

Ein grundsätzliches Problem stellt die Behandlung der direkten und indirekten Rede dar. Die wörtliche (direkte) Rede setzt sich formal aus zwei Teilen zusammen, einem übergeordneten Hauptsatz mit redeeinleitendem finiten Verb des Sagens oder Meinens und der untergeordneten wörtlichen Redewiedergabe. Die Redeeinführung und die Redewiedergabe können dabei sowohl durch einen Doppelpunkt getrennt als auch durch ein Komma verbunden sein. Obwohl die Redewiedergabe nur Satzgliedstatus hat und von dem redeeinleitenden Hauptsatz anhängt, werden dennoch beide Teile als voneinander unabhängige eigenständige Sätze betrachtet. Dies ist deshalb von Vorteil, weil die Entscheidung somit eine einheitliche Behandlung der wörtlichen Rede zuläßt. Besteht die wörtliche Rede aus mehreren Abschnitten, die durch satzabschließende Interpunktionszeichen voneinander getrennt sind, so werden diese nach den geltenden Regelungen als eigenständige Sätze aufgefaßt.

Da nicht alle Autoren die wörtliche Rede gleich realisieren, soll anhand einiger Beispiele aufgezeigt werden, welche Vorkommensweisen beobachtet wurden und wie sie zu bewerten sind. Zur besseren Übersicht werden dabei jeweils die Teile eingeklammert, die als ein Satz aufzufassen sind.

(1) Sowohl der redebezeichnende Satz als auch die Redewiedergabe sind syntaktisch vollständig und durch ein satzabschließendes Zeichen als eigenständige Sätze gekennzeichnet. Die wörtliche Rede ist durch Anführungszeichen markiert.

Beispiel: *[Nach einer Weile sagte er folgendes:] [„Du hast recht!"]* = 2 Sätze (1 Clause / 1 Clause)

(2) Der erste Fall tritt fast nie auf. Für gewöhnlich fehlt im redebezeichnende Satz das formale Objekt. Es kann jedoch ergänzt werden, wodurch sich dann wieder zwei vollständige Sätze ergeben..

Beispiel: *[Er sagte:] [„Die Wölfe kommen zurück.] [Sie wittern den Frieden."]* = 3 Sätze (1 Clause / 1 Clause / 1 Clause)

(3) Der redebezeichnende Satz ist der Redewiedergabe nachgestellt und nur durch ein Komma abgetrennt. Die beiden Teile der wörtliche Rede lassen sich vertauschen, und das Komma kann durch einen Doppelpunkt ersetzt werden.

Beispiel: *[„Du siehst elend aus"], [sagte die Frau.]* = 2 Sätze (1 Clause / 1 Clause)

(4) Die wörtlich Rede ist nicht durch Anführungsstriche markiert, dennoch läßt sich die Struktur durch die Redeeinführung sowie das finite Verb der Redewiedergabe im Indikativ als wörtliche Rede erkennen und wird entsprechend bewertet.

Beispiel: *[Mir steht's bis hier, das kann ich dir sagen], [sagte der Ziska.]* = 2 Sätze (2 Clauses / 1 Clause)

(5) In Sätzen, die außer der wörtlichen Rede weitere Bestandteile aufweisen, werden gleichgeordnete Satzteile als ein Satz aufgefaßt.

Beispiel: *[„Das darfst du doch nicht"], [sagte sie lediglich und sah sich dann die Leibold und den Leibold auf eine recht genüßliche Weise an.]* = 2 Sätze (1 Clause / 2 Clauses)

(6) Die Regelung für die wörtliche Rede gilt auch dann, wenn der Autor dem stilistischen Prinzip *variatio delectat* folgt und der redeeinführende Ausdruck ein finites Verb aufweist, das zur Realisierung nicht zwingend einer Äußerung bedarf oder nur die Begründung bestimmter Haltungen oder Stimmungen ist (vgl. Cherubim, 1993:32).

Beispiel: *[„Die gesamte deutsche Wirtschaft transferiert zur Zeit ihr Geld ins Ausland"], [beobachtet ein Luxemburger Vermögensverwalter.]* = 2 Sätze (1 Clause / 1 Clause)

Während bei der direkten Rede immer eine Aufsplittung in zwei getrennte Sätze vorgenommen wird, liegt der Fall der indirekten Rede etwas anders. Da bei der referierenden Rede nie eine Abtrennung zwischen dem einleitenden Satz und der Redewiedergabe durch abschließende Satzzeichen stattfindet, werden Sätze mit indirekter Rede stets als *ein* Satz aufgefaßt. Indirekte Rede liegt vor, wenn von einem Verb des Sagens, Meinens oder Denkens ein Gliedsatz mit einleitender Konjunktion (ob, daß) abhängig ist bzw. dem Hauptsatz ein uneingeleiteter Nebensatz folgt, dessen finite Verbformen als Konjunktiv 1 oder in Ausnahmefällen als Konjunktiv II gekennzeichnet sind.

Beispiele: *In diesem Sommer schließlich entschied der Bundesgerichtshof, die Bank habe sich nicht „anlegergerecht" verhalten und müsse dem Ehepaar das Geld erstatten.* = 1 Satz (3 Clauses)
Der Leibold behauptete wie alle Kunden, daß sie zu spät kämen. = 1 Satz (2 Clauses)

Im Zusammenhang mit der direkten und indirekten Rede traten noch zwei weitere Fälle auf, die ebenfalls betrachtet werden müssen:

(7) Im SPIEGEL findet bei einzelnen Sätzen eine Vermischung zwischen direkter und indirekter Rede statt. Es wird nach den geltenden Regeln der redeeinleitende Teil zusammen mit der indirekten Rede als ein Satz und die wörtliche Redewiedergabe als ein weiterer Satz aufgefaßt.

Beipiel: [„ Jeder Krieg beschädigt das Erbgut eines Volkes"], [sagt er düster, und auf dem Balkan sei ein dauerhafter Friede

(8) In fachwissenschaftlichen und philosophischen Aufsätzen werden in einigen Fällen innerhalb eines Satzes Teile der Argumentationsstruktur von anderen Autoren übernommen, was meistens durch ein „wie. ..sagte" angezeigt wird. Analog zu den getroffenen Vereinbarungen werden solche Konstruktionen als eigenständige eigenständige Sätze gewertet.

Beispiel: Aus der unbezweifelbaren Tatsache, daß die Urteilsfähigkeit, [wie Mathieu sagt], abgedankt hat, darf nicht der falsche Schluß gezogen werden, als habe das Handwerk abgedankt. = 2 Sätze, 3 Clauses, 1 Clause

2.3.2 Parenthese

Eine Parenthese ist eine „Phrase (Einzelwort oder Wortgruppe) oder Satzkonstruktion, die in einen Satz eingeschoben ist, ohne eine engere syntaktische Verbindung mit ihm einzugehen" (vgl. Kürschner, 1993:227). Die Trennung der Parenthese vom übrigen Satz erfolgt durch Kommata, Gedankenstriche oder Klammern. Parenthesen werden als zum Satz gehörend gewertet, da sie zwar „formalgrammatisch nicht in den Satz eingegliedert sind, jedoch semantisch oder psychisch motiviert mit dem ‚einbettenden' Satz zusammenhängen" (vgl. Lewandowski, 1990).

Beispiel: *Man mag argumentieren, daß der Kommunismus für die Entwicklungsländer wie Rußland einen Modernisierungsschub gebracht hat - ich habe das für den Nationalsozialismus und Deutschland in einer umstrittenen These behauptet-, aber in den entwikkelten Ländern Ostmitteleuropas war er pure Gewaltherrschaft.* = 1 Satz (4 Clauses)

2.3.3 Ellipse

Bei Texten mit einem hohen Anteil wörtlicher Rede kann es vorkommen, daß einige Sätze kein finites Verb beinhalten. Um Häufigkeitsverteilungen zu vermeiden, die sehr viele nullclausige Sätzen aufweisen, wird der Begriff der Ellipse eingeführt. Unter Ellipse oder Reduktion versteht man die Aussparung sprachlicher Elemente, die aufgrund syntaktischer Regeln bzw. lexikalischer Eigenschaften notwendig sind. (vgl. Bußmann, 1990:207). Eine solche elliptische Verkürzung ist möglich, wenn der inhaltliche Zusammenhang soweit gesichert ist, daß eine Wiederholung identischen Materials unterbleiben kann oder wenn eine bestimmte Struktur in vergleichbaren Situationen immer wieder begegnet (vgl. Brinkmann, 1974:144). Verschiedene sprachliche Konstruktionen lassen sich als Ellipse auffassen. Im Zusammenhang der Arbeit interessieren jedoch nur solche, die finite Verbformen aussparen, da alle übrigen Reduktionen keine Auswirkung auf die Satzlänge haben. Zu betrachten sind folgende Fälle:

Prädikatellipse

Auf die Wiederholung identischer Prädikate innerhalb eines Satzes wird verzichtet. Eine elliptische Aussparung wird jedoch nur dann angenommen, wenn sich das Prädikat auf unterschiedliche Satzglieder in der Funktion von Subjekten bezieht, die nicht gemeinsam ersetzbar sind bzw. im Aussagesatz nicht zusammen die Strukturposition vor dem finiten Verb einnehmen können. Keine Ellipse wird angenommen, wenn das finite Verb mehrere Partizipien oder Infinitive koordiniert.

Beispiele:

Die westlichen Demokratien haben manchmal Unrecht und ihre Gegner Recht. = *Die westlichen Demokratien haben manchmal Unrecht und ihre Gegner haben Recht.* = 2 Clauses (Ellipse)
Sie können die Wertpapiere ja aufteilen und zweimal vorbeikommen. = 1 Clause (Koordination)

Frage-Antwort-Paare

Im Gespräch bauen die Beteiligten eine gemeinsame Satzfolge auf. Vor allem in Frage-Antwort-Sequenzen sind die Strukturen der Frage und des Antwortsatzes äquivalent, so daß die Antwort auf notwendige Elemente verkürzt wird.

Beispiel: „Kannst Du noch?" fragte der Mann. *„Nein!"* = *„Nein, ich kann nicht mehr."* = 1 Clause

Satzwertige Aussagen

Satzwertige Aussagen wie *Bitte, Danke, Guten Tag, Auf Wiedersehen*, die ohne finites Verb gebraucht werden, werden ebenfalls als elliptische Sätze aufgefaßt und entsprechend ausgewertet.

Adjektivsätze

Adjektivsätze sind häufig reduziert, wenn sie eine Reaktion auf ein Verhalten oder Geschehen formulieren. Je nach Kontext kann die Reduktion so weit führen, daß das Adjektiv zur satzwertigen Aussage wird.

Beispiele:

Schön, daß du gekommen bist. = *Es ist schön, daß du gekommen bist.* = 2 Clauses
Sehr witzig! = Das findest du wohl sehr witzig! = 1 Clause

2.3.4 Satzwertiges Partizip und satzwertiger Infinitiv

Da die Definition der Satzlänge nur finite Verbformen zuläßt und Partizipien sowie Infinitive demnach nicht bewertet werden, müssen satzwertige Partizipien und satzwertige Infinitive als Sonderfälle geklärt werden. Handelt es sich um eine Konstruktion mit satzwertigem Partizip II, so wird das satzwertige Partizip als ein Clause gezählt, da in solchen Fällen lediglich das finite Verb getilgt ist.

Im Gegensatz zum Partizip II treten bei satzwertigem Partizip 1 bzw. satzwertigem Infinitiv nie finite Verbformen auf, weshalb diese Konstruktionen als nullwertig aufgefaßt werden. Zwar lassen sich solche Strukturen durch vollständige Nebensätze paraphrasieren, derartige syntaktische Transformationen werden aber vermieden, um die Oberflächenstruktur des Satzes nicht zu stark zu verändern.

Beispiele:

Das Dorf lag in der Mitte offener Felder, rundum von Wäldern umstellt. = *Das Dorf lag in der Mitte offener Felder und war rundum von Wäldern umstellt.* = 2 Clauses
Die bisher erörterten Integrationswege sind bestenfalls geeignet, Aspekte ein- es politischen Integrationsprozesses anzudeuten und künftige europäische Integrationsstrategien anzureichern. = 1 Clause

2.3.5 Herausstellungen

Der Terminus *Herausstellung* kennzeichnet im folgenden sowohl Rechts- als auch Linksversetzungen.

Der wichtigste Fall der Herausstellung nach rechts ist die Apposition, die sich in der Regel durch Kasuskongruenz mit ihrem Bezugswort auszeichnet, aber auch im Nominativ stehen kann. Da Appositionen keine finiten Verbformen aufweisen und der Text möglichst unverändert betrachtet werden soll, werden solche Fügungen nicht als Clauses aufgefaßt. Analog zu dieser Festsetzung werden alle übrigen Herausstellungen nach rechts ebenfalls nicht als Clauses angesehen, zumal sie meistens die Funktion spezifizierender Zusätze haben und somit ohnehin „appositionsverdächtig" sind (vgl. Schindler, 1990:334).

Linksversetzungen, die nur vereinzelt in den Texten auftreten, werden ebenfalls nicht als Clauses gewertet, wodurch sich eine einheitliche Regelung zur Behandlung von Herausstellungen ergibt.

Beispiele:

John Alcrock, Chef der Niederlassung der BHF-Bank aus Jersey, schaut ganz konsterniert, wenn er auf ein mögliches Steuerdiktat aus Brüssel angesprochen wird. = 1 Satz (2 Clauses) (Apposition)
Sie hat eine Wunde am Bein gehabt, am rechten Vorderbein. = 1 Satz (1 Clause) (Nachtrag)

Seine Uhr, Tom, seine Uhr geht heute nach. = 1 Satz (1 Clause) (Wiederholung eines Wortes)

Probleme bei der Behandlung der Nachträge ergeben sich in einigen Fällen dadurch, daß diese durch ein satzabschließendes Zeichen vom übrigen Satz getrennt ist. Da der Punkt bzw. der Doppelpunkt in solchen Fällen nur eine betonende Funktion hat, wird er nicht als satzabschließend gewertet. Der Nachtrag wird somit nicht als eigenständiger Satz aufgefaßt, sondern im Zusammenhang mit der vorangehenden Struktur als ein Satz aufgefaßt.

Beispiel: *Mein Vater ist jetzt auch dabei. Deinetwegen!* = 1 Satz (1 Clause)

2.3.6 Klammern

Eingeklammerte Ausdrücke werden genau dann bei der Auswertung mitgezählt, wenn es sich um Erklärungen oder nähere Erläuterungen handelt, die in den laufenden Text integriert werden können. Unberücksichtigt bleiben dagegen Klammerterme mit verweisen der Funktion.

Beispiele:

Doch gemäß der Rechtsprechung müssen sich bislang nur die Opfer ärztlicher Übergriffe, die mißhandelten Frauen, begutachten lassen (und dies findet dann auch noch in der geschlossenen Abteilung eines psychiatrische Landeskrankenhauses statt). = 1 Satz (2 Clauses)
Im SPIEGEL -Interview (siehe Seite 63) fordert er ein radikales Vorgehen des Gesetzgebers. = 1 Satz (1 Clause)

2.3.7 Abkürzungen

Abkürzungen werden stets in ihre ungekürzte Schreibweise umgeformt (z.B. = zum Beispiel) und, sofern sie eine finite Verbform enthalten, als ein Clause gezählt. Eingeklammerte Abkürzungen mit verweisen der Funktion bleiben gemäß der Regelung unter 4.3.6 unberücksichtigt.

Beispiele:

Der Störung entspricht die Entstörung, d.h. die Zurücknahme der Zuwendung zu sich selbst. = Der Störung entspricht die Entstörung, das heißt die Zurücknahme der Zuwendung zu sich selbst. = 1 Satz (2 Clauses)
Insbesondere konnte der Vormund den Aufenthalt des Mündels bestimmen und notfalls dieses mit Genehmigung des Vormundschaftsgerichtes in eine geschlossene Anstalt (Landeskrankenhaus) überstellen (vgl. §§ 1800, 1631 b BGB mit §§ 64a ff FGG alter Fassung). = 1 Satz (1 Clause)

2.3.8 Fremdsprachiges

Das Problem fremdsprachiger Ausdrücke im Satz läßt sich auf zwei generelle Fälle beschränken. Fremdwörter und Phrasen ohne finites Verb können als Bestandteil des Satzes akzeptiert werden, da sie keine Auswirkung auf die Satzlänge haben. Dagegen werden fremdsprachige syntaktische Konstruktionen mit finitem Verb sowie fremdsprachige Sätze aus dem Textbestand ausgesondert, wobei für den ersten Fall zusätzlich anzumerken ist, daß nur der fremdsprachige Teil getilgt werden muß, während der restliche Satz nach den geltenden Regeln ausgewertet wird.

Beispiele:

Natura parendo vincitur. Der gesamte Satz ist nicht in die Unterschung eingegangen.
Ein gutes Beispiel liefert die Formel „Black is beautiful", während ich die Formel „White is beautiful" noch nicht gehört habe; sie wäre ebenso falsch, aber weniger dynamisch. = 3 Clauses

2.4 Das Verfahren

Bei der Auswertung des Textmaterials findet ausschließlich der laufende Text Beachtung, weder Überschriften, Untertitel noch der Verfasser gehen mit ein. Ebenso findet keine Auszählung der Textbestandteile in Tabellen, Bildunterschriften und Fußnoten statt; auch Gliederungsmarkierungen, zumeist kursiv gedruckt, bleiben unbeachtet. Widmungen und dem Text voran gestellte Zitate, wie sie in einigen fachwissenschaftlichen und philosophischen Texten zu finden sind, werden ebenfalls nicht ausgewertet. Für die Artikel aus dem SPIEGEL gilt insbesondere, daß der den Artikel zusammenfassende fettgedruckte Text unterhalb des Titels bei der Auszählung nicht berücksichtigt wird, da davon auszugehen ist, daß solche Textbestandteile nachgetragen worden sind und somit eine andere Struktur aufweisen als der übrige Text.

Die Datenerhebung erfolgt so, daß jeweils alle Sätze eines Textes nacheinander ausgezählt und keine Stichproben erhoben werden, denn ein Text ist eine geschlossene Einheit und „enthält möglicherweise Längenzyklen oder Längenmuster, oder sogar Abhängigkeiten zwischen den Nachbarsätzen, die zur Längengestaltung beigetragen haben" (Köhler & Altmann, Manuskript:6). Die Untersuchung beschränkt sich auf die Auswertung der Einzeltexte. Es werden keine Zusammenfassungen mehrerer unterschiedlicher Texte vorgenommen, da durch die unterschiedlichen Rhythmen und die Überdeckung verschiedener Stile Inhomogenitäten zu erwarten sind, die zu Verschiebungen bzw. zur Verwerfung des Modells führen können. Bei solch zusammengesetzten Verteilungen müßte der besonderen Struktur durch entsprechende zusammengesetzte theoretische Verteilungen Rechnung getragen werden (vgl. Altmann, 1988:156).

Eine Fehlerquelle, die diesem Verfahren anhaftet, liegt in der Art und Weise der Datenerhebung, denn die Auszählung der Satzlängen erfolgte mit Hilfe von Strichlisten. Durch manuelle Eintragungen ist es möglich, irrtümlich einen Strich in einer falschen Spalte zu machen. Durch mehrmaliges Auswerten eines jeden Textes und den anschließenden Vergleich der Ergebnisse sollte der Fehler mini-

miert werden. Dennoch sind Fehler nicht ganz auszuschließen, und die Fehlerquote liegt bei etwa einem Prozent. Für künftige Untersuchungen ist zu überlegen, ob man anders operationalisiert, um dann eine maschinelle Auswertung mittels Computer vornehmen zu können, in der Hoffnung, daß sich so der Fehler noch weiter begrenzen läßt. Ausgewertet wurden die folgenden 85 Texte:

Text 1: Ole Schultheis, Zirkus auf dem Bauernhof
Text 2: Margret Rettich, Geschichte ohne Ende
Text 3: Margret Rettich, Trines Spuk
Text 4: Margret Rettich, Die Landstraßengeschichte
Text 5: Margret Rettich, Michels Kaninchen
Text 6: Ursula Wölfel, Das Miststück
Text 7: Ursula Wölfel, In einem solchen Land
Text 8: Rolf Krenzer, Die Sache mit der Schultasche
Text 9: Achim Bröger, Moritz tauscht
Text 10: Achim Bröger, Moritz lernt das Schwimmen
Text 11: Peter Härtling, Der Ausreißer
Text 12: Gert Prokop, Die Maus im Fenster
Text 13: Elisabeth Borchers, Reise mit Samuel
Text 14: Margarete Lengauer, Geschichte einer Großmutter
Text 15: Klaus Kordon, Zahl oder Adler
Text 16: Gudrun Pausewang, Sascha und Elisabeth
Text 17: Monika Pelz, Der Wind in der Krottenbachstraße
Text 18: Christa Reinig, Die Wölfin
Text 19: Urs Widmer, Tod und Sehnsucht
Text 20: Ilse Aichinger, Engel in der Nacht
Text 21: Ilse Aichinger, Die geöffnete Order
Text 22: Martin Walser, Gefahrenvoller Aufenthalt
Text 23: Martin Walser, Der Umzug
Text 24: Gerd Gaiser, Der Schlangenkönig
Text 25: Luise Rinser, Die rote Katze
Text 26: Hans Bender, Mein Onkel aus Amerika
Text 27: Hans Bender, Die Wölfe kommen zurück
Text 28: Hans Bender, In der Gondel
Text 29: Wolfdietrich Schnurre, Auf der Flucht
Text 30: Gisela Elsner, Die Mieterhöhung
Text 31: Gabriele Wohmann, Wiedersehen in Venedig
Text 32: Siegfried Lenz, Die Flut ist pünktlich
Text 33: Siegfried Lenz, Ein Haus aus lauter Liebe
Text 34: Heinrich Böll, Der Mann mit den Messern
Text 35: Service nur für Steuersünder
Text 36: Ingolf Doler, „Klagt nicht an der Mauer"
Text 37: Hans Magnus Enzensberger, Ausblicke auf den Bürgerkrieg
Text 38: Gisela Friedrichsen, Hört das nie auf?
Text 39: Walter Mayr, „Sie sind zäh wie die Teufel"
Text 40: Walter Mayr, Im Zahnrad der Zeit
Text 41: Walter Mayr, Totentanz im Ghetto
Text 42: Joachim Preuss, Der gedemütigte Held

Text 43: Waffen für den Todfeind
Text 44: X für unbekannt
Text 45: Die Bank gewinnt immer
Text 46: Rudolf Augstein, „Freunde für immer"
Text 47: Kämpfen und Kungeln
Text 48: Recycling ist nur der zweitbeste Weg
Text 49: Wolf Lepenies, Vorwärts mit der Aufklärung
Text 50: Bernd Dörler, „Unser Marsch hat begonnen"
Text 51: Ralf Dahrendorf, Eine große, universelle Sicht
Text 52: Prof. Dr. Hans J. Friedrichs, Das neue Betreuungsgesetz
Text 53: Wiss. Mitarbeiter Frank L. Lorenz, Beschlagnahme von Krankenunterlagen - Prozessuale Anmerkungen zur Memmingen-Entscheidung des BGH
Text 54: Prof. Dr. Dieter Strauch, Rechtsgrundlagen der Haftung bei Rat, Auskunft und Gutachten
Text 55: Privatdozent Dr. Dieter Dörr, Der „numerus clausus" und die Kapazitätskontrolle durch die Verwaltungsgerichte
Text 56. Carsten Peter und Ralf Ludwig, Das Grund recht auf Kriegsdienstverweigerung
Text 57: Prof. Dr. Edgar Ruhwedel, Grundlagen und Rechtswirkungen sogenannter relativer Verfügungsgebote
Text 58: Prof. Dr. Rolf Hofmann, Zusammenarbeit zwischen interner Revision und Abschlußprüfung - Instrumente zur Erhöhung des Wirkungsgrades
Text 59: Prof. Dr. Laurenz Lachnit, Erfolgs- und Finanzplanung für mittelständische Betriebe als Electronic-Banking-Leistung der Kreditinstitute
Text 60. Dr. Carsten T. Jebens, Rückstellungen auf schwebende Waren - Einkaufskontrakte für einen unterdurchschnittlichen Unternehmergewinn
Text 61: Dr. Winfried Paschek, Chancen und Forderungen für die rückgedeckte Gruppen-Unterstützungskasse
Text 62: Prof. Dr. Franz W. Wagner, Perspektiven der Steuerberatung: Steuerrechtspflege oder Planung der Steuervermeidung?
Text 63: Dr. Günter Vahl, Die Stellungnahme des Institutes der Wirtschaftsprüfer zur Unternehmensbewertung
Text 64: Erich Röper, Die Verfassung des Deutschen Bundes
Text 65: Peter Steinbach, Perspektiven politischer Integration
Text 66: Karl Jordan, Die Gestalt Heinrich des Löwen im Wandel des Geschichtsbildes
Text 67: Karl Dietrich Erdmann, Das Grundgesetz in der Verfassungsgeschichte
Text 68: Hartmut Bookmann, Die Vergangenheit des Deutschen Ordens im Dienste der Gegenwart
Text 69: Carl Friedrich von Weizsäcker, Kirchenlehre und Weltverständnis
Text 70: Carl Friedrich von Weizsäcker, Wozu Meditation?
Text 71: Carl Friedrich von Weizsäcker, Verteidigung der Freiheit
Text 72: Karl Popper, Selbstbefreiung durch das Wissen
Text 73: Karl Jaspers, Die Philosophie in der Welt
Text 74: Karl Jaspers, Die Aufgabe der Philosophie in der Gegenwart
Text 75: Karl Jaspers, Die Idee des Arztes
Text 76: Hans-Georg Gadamer, Philosophische Bemerkungen zum Problem der Intelligenz
Text 77: Hans-Georg Gadamer, Der Tod als Frage
Text 78: Hans-Georg Gadamer, Apologie der Heilkunst
Text 79: Arnold Gehlen, Die gewaltlose Lenkung

Text 80: Helmuth Plessner, Die Musikalität der Sinne. Zur Geschichte eines modernen Phänomens
Text 81: Helmuth Plessner, Das Lächeln
Text 82: Helmuth Plessner, Der Weg der Soziologie in Deutschland
Text 83: Max Horkheimer, Soziologie und Philosophie
Text 84: Max Horkheimer, Theismus und Atheismus
Text 85: Max Horkheimer, Der Mensch in der Wandlung seit der Jahrhundertwende

3 Versuch einer mathematischen Modellierung

Nachdem Kapitel 2 der praktischen Ermittlung der Daten gewidmet war, soll im folgenden eine Auseinandersetzung mit dem theoretischen Hintergrund der Untersuchung stattfinden.

3.1 Die 0-gestutzte negative Binomialverteilung

Schon vor dem Beginn des Schreibprozesses bestimmen die Wahl des Themas und der Gattung, der intendierte Hörerkreis, die geplante Textlänge und -struktur die Gestaltung eines Textes. Bei der eigentlichen Textproduktion entwickelt sich dann nach und nach eine geordnete Folge von den Text konstituierenden Elementen, indem der Autor nach seiner Wahl Sätze kombiniert, so daß sie einen Sinnzusammenhang ergeben. Bei der Erstellung von Prosa- oder Sachtexten unterliegen neben der korrekten grammatischen Struktur vor allem inhaltliche Aspekte der Kontrolle des Autors. Als unbewußter Prozeß dagegen erweist sich die Gestaltung der Satzlänge. Solange der Text nicht einem bestimmten Rhythmus unterliegen soll, hat der Autor nichts weniger im Sinn, als die Länge seiner Sätze zu kontrollieren. Dies führt dazu, daß sowohl sehr kurze Sätze mit nur einem einsilbigen Wort (z.B. *Sprich!*) als auch extrem lange Sätze auftreten können. Da man von jedem Satz durch Hinzufügen eines weiteren Teilsatzes eine noch längere Form bilden kann, sind theoretisch unendlich lange Sätze denkbar. Praktisch ist diese Möglichkeit jedoch durch die Forderung des Lesers nach nicht zu aufwendiger Dekodierung des Textes eingeschränkt, und die Zahl der Clauses pro Satz erweist sich auf der Ebene des konkreten Textes stets als endlich.

Obwohl der Autor der Länge seiner Sätze keine Beachtung schenkt, ergeben sich bei der Auszählung der Längen in geschlossenen Texten dennoch „anständige" Häufigkeitsverteilungen (vgl. Köhler & Altmann, Manuskript:7). Geht man davon aus, daß Sprache ein selbstregulierendes System ist, so läßt sich ihr Zustandekommen dadurch erklären, „daß sich die Länge im Laufe der Texterzeugung selbst organisiert, wobei sie die Proportionalitätsbedingung im Sinne der Rahmenbedingungen interpretiert" (Köhler & Altmann, Manuskript:8). Es gilt demnach die Proportionalitätsvorschrift $P_x \sim P_{x-1}$ und deshalb $P_x = g(x)\, P_{x-1}$, wo-

bei $P_x = P(X = x)$ die Wahrscheinlichkeit ist, daß die Satzlänge X den Wert x annimmt, und $g(x)$ eine Funktion bezeichnet, die das Verhältnis der benachbarten Wahrscheinlichkeiten regelt.

Wie aber sieht die Funktion $g(x)$ aus? „Die Bedürfnisse des Menschen gestalten die Sprache so, daß sie das Verhalten eines zweckgebundenen (purposeful) Systems annimmt" (Köhler & Altmann, 1986:258) d. h. also, daß Sprache sich nicht zufällig gestaltet, sondern gewissen Bedingungen unterlegen ist. Um zu einem passenden Modell zu gelangen, müssen diese Bedingungen bereits im Ansatz berücksichtigt werden. Neben dem Autor, der die Freiheit der Gestaltung hat und den Text vor allem durch seinen Stil prägt, wirkt auch der Hörer, dem die Dekodierung des Textes obliegt, indirekt auf diesen ein, damit die Kommunikation reibungslos verlaufen kann. Schließlich beeinflußt noch das Textvorbild den aktuellen Text, so daß man mindestens zu drei textbestimmenden Faktoren gelangt: Textvorbild, Wirkung des Autors (Sprecher), Wirkung des Lesers (Hörer).

Die Wirkung a des Textes ist konstant, b und c dagegen sind Funktionen von x. wobei c bremsend auf b wirkt, um die Gestaltungsmöglichkeiten des Autors einzuschränken und den Erhalt der Sprache zu gewährleisten (vgl. Altmann, 1988:152). Es gilt $b < c$ wodurch praktisch die Konvergenz der Verteilung gewährleistet ist. Man geht davon aus, daß sich mit Hilfe der drei genannten Faktoren $g(x)$ als gebrochen rationale Funktion der folgenden Form darstellen läßt

$$g(x) = \frac{a+bx}{cx}, \quad x \neq 0$$

und erhält damit die Gleichung

$$P_x = \frac{a + bx}{cx} P_{x-1}$$

Aufgrund der Festsetzung, daß auch elliptisch vorhandene finite Verben berücksichtigt werden, existieren keine Sätze der Länge $x = 0$. Die kleinste Länge, die ein Satz annehmen kann, ist demnach $x = 1$. Da Köhler & Altmann für diesen Fall nur das Ergebnis angeben, soll im folgenden die Herleitung der 0-gestutzten negativen Binomialverteilung formuliert werden, auch wenn sie analog zum ungestutzten Fall erfolgt (vgl. auch Köhler & Altmann, Manuskript:8ff; Altmann & Best, 1996). Ausgehend von der Gleichung

$$P_x = \frac{a + bx}{cx} P_{x-1}$$

erhält man durch rekursives Einsetzen:

$$P_x = \frac{a+xb}{xc} P_{x-1} = \frac{(a+xb)}{xc} \frac{[a+(x-1)b]}{(x-1)c} P_{x-2} = \ldots$$

$$= \frac{(a+xb)[a+(x-1)b]\ldots(a+3b)(a+2b)}{x!c^{x-1}} P_1$$

$$= \frac{1}{x!\ c^{x-1}} \left[\prod_{l=2}^{x} (a+lb) \right] P_1$$

$$= \frac{1}{x!\ c^{x-1}} \left[\prod_{l=2}^{x} b\left(\frac{a}{b}+l\right) \right] P_1$$

$$= \frac{b^{x-1}}{x!\ c^{x-1}} \left[\prod_{l=2}^{x} \left(\frac{a}{b}+l\right) \right] P_1$$

Definiert man nun $q := \frac{b}{c}$ mit $0 < q < 1$ und $r := \frac{a}{b}$, so ergibt sich für P_x:[6]

$$P_x = \frac{q^{x-1}}{x!} \left[\prod_{l=2}^{x} (r+l) \right] P_1 .$$

Aus der Gleichheit

$$\frac{1}{x!} \prod_{l=2}^{x} (r+l) = \frac{1}{x!} \left[\prod_{l=2}^{x} (r+l) \right] \frac{(r+1)!}{(r+1)!} = \frac{(r+x)!}{x!(r+1)!} = \binom{r+x}{x} \frac{1}{r+1}$$

folgt, daß sich P_x schreiben läßt als:

$$P_x = \binom{r+x}{x} \frac{1}{r+1} q^{x-1} P_1 .$$

Im folgenden Schritt setze man nun $k := r + 1$. Man erhält dann:

$$P_x = \binom{k+x-1}{x} \frac{1}{k} q^{x-1} P_1 .$$

Es bleibt schließlich nur noch P_1 zu bestimmen. Aus der Tatsache, daß die Summe der Wahrscheinlichkeiten 1 ist

$$\sum_{x=1}^{\infty} \binom{k+x-1}{x} \frac{1}{k} q^{x-1} P_1 = 1$$

und die Gleichungen

[6] $0 < q < 1$ folgt aus der Tatsache, daß $b < c$ gilt.

$$\binom{k+x-1}{x} = (-1)^x \frac{(-k)(-k-1)\ldots(-k-x+1)}{x!} = (-1)^x \binom{-k}{x} .$$

$$\sum_{x=0}^{\infty} \binom{-k}{x} (-q)^x = (1-q)^{-k}$$

gelten, berechnet sich P_1 mit $p := 1 - q$ zu:

$$1 = \sum_{x=1}^{\infty} (-1)^x \binom{-k}{x} q^{x-1} \frac{1}{k} P_1$$

$$= -\sum_{x=1}^{\infty} (-q)^{x-1} \binom{-k}{x} \frac{1}{k} P_1$$

$$= -\sum_{x=0}^{\infty} (-q)^{x-1} \binom{-k}{x} \frac{1}{k} P_1 + (-q)^{-1} \frac{1}{k} P_1$$

$$= \frac{1}{q} \sum_{x=0}^{\infty} (-q)^x \binom{-k}{x} \frac{1}{k} P_1 - \frac{1}{q} \frac{1}{k} P_1$$

$$= \frac{1}{q} \frac{1}{k} P_1 (p^{-k} - 1).$$

Durch Umformen dieser Gleichung ergibt sich:

$$P_1 = \frac{qkp^k}{1-p^k} .$$

Einsetzen von P_1 in P_x liefert schließlich die negative Binomialverteilung in ihrer 0-gestutzten Form:

$$P_x = \binom{k+x-1}{x} \frac{1}{k} q^{x-1} \frac{qkp^k}{1-p^k}$$

$$= \binom{k+x-1}{x} \frac{q^x p^k}{1-p^k}, \quad x = 1, 2, \ldots$$

Für $k \to \infty$, $q \to 0$ und $kq \to a$ läßt sich die 0-gestutzte negative Binomialverteilung durch die positive Poissonverteilung

$$P_x = \frac{e^{-a} a^x}{x!(1-e^{-a})}, \quad x = 1, 2, \ldots, a > 0$$

approximieren.

Die iterative Berechnung der theoretischen Verteilungen sowie die Tests erfolgten mit Hilfe des Programms Altmann-Fitter.

3.2 Die Schätzparameter *p* und *k*

Zur Überprüfung des Modells der 0-gestutzten negativen Binomialverteilung benötigt man Werte für die Parameter p und k. Da bislang noch keine theoretischen Aussagen gemacht werden können, inwieweit sich die Parameter aus anderen Gesetzen ableiten lassen, bleibt zunächst nur die Möglichkeit, sie aus den vorhandenen Daten zu schätzen, was im allgemeinen mit Hilfe der entsprechenden Anwendungssoftware geschieht. Hat man jedoch kein passendes Programm zur Hand, so können auch einfache Schätzverfahren angewendet werden, die in der Regel ebenfalls brauchbare Schätzer liefern:[7]

Im folgenden bezeichne x wieder die Satzlänge und n_x die Anzahl aller Sätze eines Textes mit der Länge x. N sei die Anzahl der Sätze pro Text. Ausgehend von den ersten zwei Momenten der gestutzten negativen Binomialverteilung ergeben sich unter Nutzung von n_1 sowie

$$\bar{x} = \frac{1}{N}\sum_x x n_x$$

und

$$s^2 = \frac{1}{N}\sum_x (x-\bar{x})^2 n_x$$

für die Schätzer $\hat{p}$ und $\hat{k}$ die folgenden Gleichungen:

$$\hat{p} = \frac{\bar{x}}{s^2}\left(1-\frac{n_1}{N}\right), \qquad \hat{k} = \frac{\hat{p}\bar{x} - n_1/N}{1-\hat{p}}.$$

$\hat{p}$ und $\hat{k}$ sind konsistente, aber nicht erwartungstreue Schätzer für p und k. Sie haben zwar den Vorteil, daß sie leicht zu berechnen sind, allerdings ist die Effizienz der benutzten Methode für kleines p nicht besonders hoch. In solchen Fällen bevorzugt man gewöhnlich die Maximum-Likelihood-Methode, deren Schätzer hier in einer modifizierten Form angegeben werden (vgl. Brass, 1958:64):

$$\hat{p} = \frac{\hat{k} + n_1/N}{\hat{k} + \bar{x}},$$

$$\frac{\bar{x}\left(\hat{k}+n_1/N\right)}{\hat{k}\left(\bar{x}+n_1/N\right)} \ln\left(\frac{\hat{k}+n_1/N}{\hat{k}+\bar{x}}\right) + \frac{1}{N}\sum_{l=1}^{R}\left(\hat{k}+l-1\right)^{-1}\sum_{j=l}^{R} n_j = 0.$$

[7] Vgl. W. Brass, 1958: 59 f. R. Köhler & G. Altmann (Manuskript) geben ebenfalls die erste Methode zu Schätzung der Parameter an (vgl. S. 14).

R bezeichnet dabei den größten Wert der Stichprobe.
Die Gleichung in $\hat{k}$ kann durch gängige Iterationsverfahren gelöst werden, wobei sich der gewonnene Schätzer der ersten Methode als Anfangswert verwenden läßt. $\hat{p}$ ergibt sich durch Einsetzen von $\hat{k}$ in die obere Gleichung. Sollten sich mittels dieser Methoden keine brauchbaren Ergebnisse erzielen lassen, so verwende man andere Schätzverfahren (z.B. Sampford, 1955). Für die Schätzer der positiven Poissonverteilung greife man auf entsprechende Methoden zurück (z.B. David & Johnson, 1952).

3.3 Der Chiquadrat-Test

Ziel der Untersuchung ist es, die 0-gestutzte negative Binomialverteilung als Modell der Satzlängenverteilung zu testen. Als Kriterium für die Güte der Anpassung wird übliche Chiquadrat-Test herangezogen. Akzeptiert werden in der vorliegenden Arbeit zwar Anpassungen mit $\alpha = 0.01$ also $P > 0.01$, gute Ergebnisse lassen sich jedoch erst mit $P > 0.05$ erzielen. Die Festsetzung dieses sehr niedrigen Niveaus $P > 0.01$ läßt sich dadurch rechtfertigen, daß es bereits als Erfolg angesehen werden muß, wenn überhaupt eine Anpassung an das Modell möglich ist. Eine weitere Begründung ergibt sich aus der Wahl des Testverfahrens. Die Auswertung der Untersuchung könnte auch mit spezielleren Testverfahren durchgeführt werden, die bessere Ergebnisse als das X^2 erzielen. „Die Kontrolle anderer Koeffizienten, wie z.B. λ^2 von Pederson & Johnson (1990), bei dem auch die Freiheitsgrade und die linearen Abweichungen in Betracht gezogen werden, verbessert den Eindruck beträchtlich. (ii) Die Anwendung anderer Anpassungstests, wie z.B. der Systeme von Cressie & Read (1984) oder Moore & Spruill (1975) u.a. zeigen, daß die Anpassungen höhere P-Werte zeigen als der klassische Chiquadrat-Test“ (Altmann & Best, Manuskript:4). Auf solche Verfahren wird in dieser Arbeit dennoch kein Bezug genommen, da nur die Verwendung allgemein bekannter Verfahren die Einheitlichkeit zwischen verschiedenen Untersuchungen gewährleisten und die Möglichkeit zum Vergleich bieten kann. Dem verhältnismäßig schlechten P-Wert des X^2-Tests wird stattdessen durch ein sehr niedriges Niveau Rechnung getragen.

Bei einigen Texten aus dem Bereich der Kurzprosa für Kinder tritt der Fall ein, daß FG = 0 gilt, d.h., die Zahl der Klassen ist so gering, daß die Berechnung keine Freiheitsgrade liefert. In solchen Fällen läßt sich P nicht berechnen, aber es kann ein anderer Maßstab für die Güte der Anpassung gefunden werden. Wir betrachten dann $C = X^2/N$. Das Ergebnis wird als gut angesehen für $C \leq 0{,}02$. Auf den C-Wert wird ebenfalls dann zurückgegriffen, wenn die Berechnung einen sehr kleinen P-Wert liefert, wodurch die Anpassung möglicherweise schon gut angesehen werden kann. Unproblematisch ist die Angabe des C-Wertes je-

doch auch nicht, denn er erweist sich zwar bei großen Stichproben als geeignet, nicht aber bei kleinen.

4 Überprüfung der Daten im Hinblick auf das gewählte Modell

Die Voraussetzung jeder Forschung ist die Feststellung und Beschreibung der zu untersuchenden Erscheinung, hier der Satzlänge. Da die Beschreibung aber nur ein erster Schritt auf dem Weg zur Erklärung ist und die quantitativen Informationen sinnvoll in Theorien eingebettet werden müssen, ist das eigentliche Ziel der Arbeit nicht die Aufstellung der Satzlängenhäufigkeitsverteilungen bei verschiedenen Texten unterschiedlicher Funktionalstile. Das Untersuchungsinteresse richtet sich vielmehr auf die Frage, ob es Modelle gibt, denen die beobachteten Häufigkeiten folgen. Speziell soll mit Hilfe des Chiquadrat-Tests geprüft werden, ob die in Kapitel 3 hergeleitete 0-gestutzte negative Binomialverteilung ein gutes Modell der Satzlängenverteilung darstellt. Mit Ausnahme von sieben Texten läßt sich die Satzlängenhäufigkeit für alle Texte mit der 0-gestutzten negativen Binomialverteilung modellieren, wodurch die Hypothese:

Die Häufigkeit der Satzlänge, gemessen in der Anzahl ihrer Clauses, folgt in ihrer Verteilung dem Modell der 0-gestutzten negativen Binomialverteilung

für gegenwartssprachliche, deutsche Texte unterschiedlicher Funktionalstile bestätigt wird. In diesem Zusammenhang läßt sich die Untersuchung auch als Beitrag zur Prüfung der allgemeineren Hypothese:

Die negative Binomialverteilung ist ein geeignetes Modell für die Längenverteilung eines Konstrukts, gemessen in der Zahl seiner Komponenten

werten. Im folgenden werden die Texte, die dem Modell folgen, einzeln aufgelistet. Anmerkung: Die mit einem Stern (*) gekennzeichneten Texte lassen sich ebenfalls durch die positive Poissonverteilung anpassen.

	Text 1		Text 2		Text 3 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	160	162.97	103	107.66	79	80.50
2	49	41.09	58	50.36	49	46.04
3	4	8.94	16	16.54	22	17.74
4	-	-	3	5.44	1	6.72
	$k = 8.8625$; $p = 0.9489$; $C = 0.0198$.		$k = 17.8170$; $p = 0.9503$; $X_1^2 = 2.46$; $P = 0.12$.		$k = 88.7772$; $p = 0.9873$; $C = 0.0020$.	

Texte 1 und 3: Da in diesen Texten $FG = 0$ gilt, können die P-Werte nicht bestimmt werden.. Es muß in diesen Fällen auf C zurückgegriffen werden.

	Text 4		Text 5 *		Text 6 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	92	93.92	104	110.00	84	87.81
2	53	48.06	69	55.87	54	45.32
3	13	16.57	13	19.02	11	15.70
4	5	4.32	4	4.88	4	5.17
5	1	1.13	1	1.23	-	-
	$k = 94.6025$; $p = 0.9893$; $X_2^2 = 1.43$; $P = 0.49$.		$k = 177.4023$; $p = 0.9943$; $X_2^2 = 5.52$; $P = 0.06$.		$k = 138.6471$; $p = 0.9926$; $X_1^2 = 3.24$; $C = 0.07$.	

	Text 7 *		Text 8		Text 9 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	130	129.45	94	94.02	119	122.38
2	35	31.96	39	38.96	47	39.47
3	2	6.29	12	12.05	6	10.15
4	1	1.08	3	3.09	-	-
5	1	0.22	1	0.88	-	-
	$k = 4.0737$; $p = 0.9027$; $X_1^2 = 3.62$; $P = 0.06$.		$k = 7.3549$; $p = 0.9008$; $X_1^2 = 0.0005$; $p = 0.98$.		$k = 148.3666$; $p = 0.9957$; $C = 0.0184$.	

	Text 10 *		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	101	104.04	125	124.26	73	70.14
2	42	34.06	52	46.72	30	35.64
3	4	8.90	14	19.45	21	18.35
4	-	-	6	8.49	9	9.50
5	-	-	7	3.81	6	4.94
6	-	-	1	1.74	3	2.57
7	-	-	1	1.53	1	1.34
8	-	-	-	-	1	1.52
	$k = 206.90$; $p = 0.9968$; $C = 0.0305$.		$k = 0.5127$; $p = 0.5029$; $X^2_4 = 6.02$; $P = 0.20$.		$k = 0.9247$; $p = 0.4720$; $X^2_5 = 1.94$; $P = 0.85$.	

Text 10: Da $FG = 0$ ist, kann P nicht bestimmt werden, so daß auf den C-Wert zurückgegriffen werden muß. Da dieser oberhalb des festgelegten Grenzwertes liegt, muß festgehalten werden, daß die Anpassung hier ein nichtsignifikantes Ergebnis liefert.

Text 11: „Er streichelt Erwin übers Haar, sagte [...], redete mit einem Zirkusfahrer, der nahm ihn am Kragen, zerrte ihn zu einem Traktor, setzte ihn neben sich, und während der ganzen Fahrt sang der Mann mit unheimlicher Fistelstimme." = 7 Clauses. Es handelt sich um eine Satzreihe, und die Länge wird ausschließlich durch die Nebenordnung verschiedener finiter Verben hervorgerufen, so daß der Satz zwar sehr lang für einen kindersprachlichen Text ist, aber dennoch nicht sehr schwierig.

Text 12: „Pipp kam, schnupperte, trippelte zum Käsekuchen, schnupperte noch einmal, probierte, fraß - und viel nicht tot um." = 7 Clauses. Die außergewöhnliche Länge für einen kindersprachlichen Text wird lediglich durch die Koordination einer Vielzahl von finiten Verben hervorgerufen und steht somit nicht im Widerspruch zu der Aussage, daß der Text auch formal den Dispositionen eines Kindes angepaßt ist.

	Text 13 *		Text 14 *		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	151	151.32	85	88.17	114	112.18
2	46	44.70	55	50.43	51	55.58
3	8	9.24	22	19.27	25	20.60
4	0	1.50	2	5.54	5	6.35
5	1	0.20	1	1.59	2	2.29
6	1	0.04	-	-	-	-
	$k = 19.2695$; $p = 0.9708$; $X^2_1 = 0.25$; $P = 0.62$.		$k = 376.5158$; $p = 0.9970$; $X^2_2 = 3.38$; $P = 0.18$.		$k = 7.1826$; $p = 0.8789$; $X^2_2 = 1.66$; $C = 0.44$	

	Text 16		Text 17 *		Text 18 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	145	145.77	80	80.84	86	89.69
2	53	52.09	47	46.02	58	50.83
3	15	14.62	18	17.59	18	19.33
4	2	3.54	6	6.55	3	5.55
5	2	0.98	-	-	2	1.60
	$k = 4.6090$; $p = 0.8725$; $X^2_1 = 0.0891$; $P = 0.77$.		$k = 143.7595$; $p = 0.9921$; $X^2_1 = 0.08$; $P = 0.77$.		k = 142.0916; $p = 0.9920$; $X^2_2 = 2.54$; $P = 0.28$.	

	Text 19 *		Text 20		Text 21 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	115	115.48	72	73.17	86	89.14
2	70	67.78	54	51.31	65	61.38
3	24	26.98	31	31.48	31	28.23
4	10	8.19	18	17.93	7	9.75
5	2	2.57	8	9.74	2	2.70
6	-	-	6	5.12	1	0.80
7	-	-	2	2.62	-	-
8	-	-	2	1.32	-	-
9	-	-	-	0.65	-	-
10	-	-	-	0.32	-	-
11	-	-	1	0.34	-	-
	$k = 55.9714$; $p = 0.9794$; $X_2^2 = 0.92$; $P = 0.63$.		$k = 2.2031$; $p = 0.5621$; $X_6^2 = 1.19$; $P = 0.98$.		$k = 526.0970$; $p = 0.9974$; $X_2^2 = 1.44$; $P = 0.49$.	

	Text 22		Text 23		Text 24	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	57	61.60	70	74.91	69	66.03
2	49	44.43	51	44.74	42	44.35
3	33	26.55	30	25.32	28	24.79
4	10	14.22	9	13.94	9	12.45
5	8	7.09	6	7.54	4	5.83
6	-	3.36	5	4.03	5	2.60
7	2	1.53	1	2.13	1	1.95
8	-	0.67	2	1.12	-	-
9	-	0.29	-	0.59	-	-
10	-	0.12	-	0.30	-	-
11	-	0.05	1	0.38	-	-
12	-	0.02	-	-	-	-
13	-	0.00	-	-	-	-
14	1	0.07	-	-	-	-
	$k = 3.1212$; $p = 0.6499$; $X_5^2 = 7.28$; $P = 0.20$.		$k = 1.3720$; $p = 0.4964$; $X_6^2 = 5.68$; $P = 0.46$.		$k = 3.0306$; $p = 0.6667$; $X_4^2 = 4.85$; $P = 0.30$.	

	Text 25		Text 26 *		Text 27	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	108	102.98	76	74.89	154	152.63
2	53	64.17	41	43.51	33	35.68
3	37	29.51	21	17.43	12	10.85
4	11	11.16	3	5.41	4	3.68
5	4	5.18	2	1.76	1	1.32
6	-	-	-	-	-	0.49
7	-	-	-	-	1	0.35
	$k = 8.3480$; $p = 0.8667$; $X_2^2 = 4.35$; $P = 0.11$.		$k = 28.2316$; $p = 0.9603$; $X_2^2 = 2.00$; $P = 0.37$.		$k = 0.0512$; $p = 0.5552$; $X_2^2 = 0.37$; $P = 0.83$.	

	Text 28		Text 29		Text 30	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	107	105.37	174	173.45	91	87.93
2	24	26.85	39	40.36	36	41.29
3	10	9.01	11	9.86	25	20.39
4	5	3.39	2	2.46	7	10.31
5	-	1.35	1	0.87	6	5.29
6	-	0.56	-	-	3	2.74
7	1	0.47	-	-	2	1.43
8	-	-	-	-	1	1.62
	$k = 0.0247$; $p = 0.5025$; $X_3^2 = 1.98$; $P = 0.37$.		$k = 0.7391$; $p = 0.7324$; $X_1^2 = 0.21$; $P = 0.65$.		$k = 0.7321$; $p = 0.4578$; $X_5^2 = 3.46$; $P = 0.63$.	

Text 30: „Die meisten Auftraggeber dieses billigen Möbeltransports, der sich nach dem Namen des Unternehmers: Millizer nannte, verbesserten sich zur Verwunderung des Ziska tatsächlich." = 1 Satz. Trotz des Doppelpunktes und der Großschreibung des folgenden Wortes wird dieser Satz als **ein** Satz aufgefaßt. Der Doppelpunkt findet hier nur Verwendung als Heraushebung eines Namens, der auch durch Kursivdruck oder Anführungszeichen gekennzeichnet werden könnte.

	Text 31 *		Text 32		Text 33	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	98	94.39	103	100.01	57	49.86
2	49	57.13	30	35.32	22	31.61
3	30	23.94	17	16.56	21	20.96
4	7	7.80	11	8.72	13	14.20
5	2	2.74	5	4.89	11	9.75
6	-	-	3	2.86	10	6.74
7	-	-	2	1.72	8	4.70
8	-	-	-	1.05	2	3.28
9	-	-	1	0.65	3	2.30
10	-	-	-	0.41	2	5.60
11	-	-	1	0.81	-	-
	$k = 24.9719$; $p = 0.9534$; $X_2^2 = 3.10$; $P = 0.21$.		$k = 0.0087$; $p = 0.2997$; $X_6^2 = 1.87$; $P = 0.93$.		$k = 0.7581$; $p = 0.2787$; $X_7^2 = 11.08$; $P = 0.14$.	

Text 33: Die beobachteten Häufigkeiten nehmen bei diesem Text bei zunehmender Länge nur sehr langsam ab. Der Text zeichnet sich durch einen hohen Anteil drei-, vier-, fünf- und sechsclausiger Sätze aus. Eine denkbare Erklärung für dieses Phänomen liefert die Tatsache, daß der Text nur wenig wörtliche Rede umfaßt und insgesamt als reflektorisch eingestuft werden kann. Um eine optimale Anpassung zu erzielen, müßte man diese Tatsache durch Modellmodifikation erfassen.

	Text 34		Text 36		Text 37	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	100	102.10	102	101.29	115	117.74
2	67	62.30	74	73.16	83	76.36
3	42	36.14	32	35.49	37	37.65
4	14	20.42	17	13.00	12	15.64
5	10	11.35	3	5.06	5	5.76
6	5	6.24	-	-	3	1.94
7	2	3.40	-	-	1	0.91
8	4	1.84	-	-	-	-
9	1	0.99	-	-	-	-
10	-	0.53	-	-	-	-
11	-	0.28	-	-	-	-
12	1	0.41	-	-	-	-
	$k = 1.3476$; $p = 0.4801$; $X_6^2 = 6.01$; $P = 0.42$.		$k = 133.5986$; $p = 0.9893$; $X_2^2 = 2.41$; $P = 0.30$.		$k = 6.1204$; $p = 0.8178$; $X_3^2 = 2.10$; $P = 0.55$.	

Text 37: Die von Enzensberger zitierten längeren Textpassagen, die sich durch den Druck vom übrigen Artikel unterscheiden, wurden nicht mitgezählt.

	Text 38		Text 40 *		Text 41 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	75	76.02	173	177.10	74	74.63
2	43	39.73	75	67.11	45	44.10
3	14	16.14	14	16.99	22	17.47
4	5	5.62	3	3.80	1	5.22
5	3	2.49	-	-	1	1.58
	$k = 5.0103$; $p = 0.8261$; $X_2^2 = 0.75$; $P = 0.69$.		$k = 446.8830$; $p = 0.9983$; $X_1^2 = 1.72$; $P = 0.19$.		$k = 177.9034$; $p = 0.9934$; $X_2^2 = 4.81$; $P = 0.09$.	

	Text 44 *		Text 45 *		Text 46 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	96	103.03	120	129.09	125	129.25
2	71	59.50	92	74.06	78	72.60
3	22	22.93	21	28.37	36	29.25
4	4	6.63	7	8.16	4	9.46
5	1	1.91	2	2.32	1	3.44
	$k = 1110.3815$; $p = 0.9990$; $X_2^2 = 4.20$; $P = 0.12$.		$k = 731.4473$; $p = 0.9984$; $X_2^2 = 7.12$; $P = 0.03$; $C = 0.0294$.		$k = 12.1616$; $p = 0.9146$; $X_2^2 = 6.97$; $P = 0.03$; $C = 0.0286$.	

Texte 45 und 46: Die P-Werte sind in diesen Texten sehr niedrig. Auch die zusätzliche Betrachtung der entsprechenden C-Werte liefert keine Ergebnisverbesserung.

	Text 47 *		Text 48		Text 49 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	143	145.19	137	142.79	74	72.87
2	54	49.99	46	41.88	60	64.47
3	11	11.51	14	11.36	42	38.16
4	-	1.99	1	2.96	18	16.99
5	1	0.32	1	0.75	7	8.51
6	-	-	-	0.18	-	-
7	-	-	-	0.04	-	-
8	-	-	-	0.01	-	-
9	-	-	-	0.00	-	-
10	-	-	-	0.00	-	-
11	-	-	-	0.00	-	-
12	-	-	1	0.03	-	-
	$k = 322.4986$; $p = 0.9979$; $X_1^2 = 1.12$; $P = 0.29$.		$k = 1.5781$; $p = 0.7725$; $X_1^2 = 1.48$; $P = 0.22$.		$k = 287.6189$; $p = 0.9939$; $X_2^2 = 1.04$; $P = 0.60$.	

Text 48: Der zwölfclausige Satz gibt die Stellungnahme verschiedener Leute zum *Grünen Punkt* an. Diese Aussagen sind jeweils nur durch einen Semikolon gegeneinander abgegrenzt.

	Text 50 *		Text 51		Text 52	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	91	96.38	110	106.93	107	110.12
2	59	49.34	54	61.08	66	60.54
3	15	16.86	30	24.15	24	25.39
4	3	5.42	6	7.42	8	8.99
5	-	-	2	2.42	3	2.83
6	-	-	-	-	-	0.81
7	-	-	-	-	-	0.22
8	-	-	-	-	1	0.10
	$k = 704.3239$; $p = 0.9985$; $X_1^2 = 3.46$; $P = 0.06$.		$k = 25.1869$; $p = 0.9564$; $X_2^2 = 2.67$; $P = 0.26$.		$k = 5.9375$; $p = 0.8415$; $X_3^2 = 0.79$; $P = 0.85$.	

Texte 52 und 52 a: In diesem Aufsatz werden die Verweise („vgl.“) auf andere Paragraphen zum Teil an die Sätze angehängt, ohne daß sie eingeklammert werden. Ist dies der Fall, so werden sie mit ausgewertet. Steht der Verweis dagegen in Klammern, bleibt er unberücksichtigt. Beispiel: „Notfalls sind sie auch jetzt wieder Vereins- und Behördenbetreuung vorgesehen, vgl. §§ 1897 ff BGB.“ = 2 Clauses.
Zu dieser Auswertung wurde noch eine alternative Zählung vorgenommen (Text 52 a), in der solche Verweise, ob eingeklammert oder nicht, grundsätzlich nicht berücksichtigt wurden. Während die erste Anpassung mit $P = 0.85$ bereits sehr gut ist, ergibt die zweite Anpassung einen nahezu optimalen P-Wert. Gleichzeitig bestätigt der Vergleich der beiden alternativen Zählungen die Festlegung, eingeklammerte Verweise nicht zu berücksichtigen. Die Auswertung 52 a zeigt ein deutlich verbessertes Ergebnis, woraus resultiert, daß eingeschobene Verweise sich störend auf den Textfluß auswirken.

	Text 52 a		Text 53 *		Text 54 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	118	117.97	88	95.88	68	68.91
2	57	56.69	74	62.33	61	56.45
3	22	22.51	27	27.05	27	32.77
4	7	7.99	7	8.81	18	15.11
5	3	2.64	1	2.93	5	5.88
6	-	0.82	-	-	2	2.01
7	-	0.25	-	-	1	0.87
8	1	0.13	-	-	-	-
	$k = 3.1810$; $p = 0.7701$; $X_3^2 = 0.09$; $P = 0.99$.		$k = 714.8701$; $p = 0.9981$; $X_2^2 = 4.47$; $P = 0.11$.		k = 14.9203; $p = 0.8971$; $X_3^2 = 2.09$; $P = 0.55$.	

	Text 55 *		Text 57 *		Text 58	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	99	105.79	83	91.37	120	119.09
2	84	75.53	84	68.12	35	36.58
3	42	35.99	27	33.88	11	10.42
4	7	12.87	15	12.65	4	3.91
5	2	3.69	1	3.78	-	-
6	1	1.13	1	1.20	-	-
	$k = 833.6704$; $p = 0.9983$; $X_3^2 = 5.86$; $P = 0.12$.		$k = 1233.2166$; $p = 0.9988$; $X_2^2 = 8.08$; $P = 0.02$; $C = 0.0369$.		$k = 1.5544$; $p = 0.7595$; $X_1^2 = 0.11$; $P = 0.74$.	

Text 57: Der P-Wert ist sehr niedrig. Auch die zusätzliche Betrachtung von C liefert keine Ergebnisverbesserung, da der Schwellenwert überschritten wird.

	Text 59 *		Text 60 *		Text 61 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	132	137.91	53	53.64	72	73.46
2	85	72.83	47	46.13	66	62.11
3	21	25.69	27	26.71	34	35.21
4	5	6.88	13	11.71	13	15.05
5	2	1.77	2	4.14	7	5.18
6	-	-	2	1.67	1	1.99
	$k = 564.4081$; $p = 0.9981$; $X_2^2 = 3.66$; $P = 0.16$.		$k = 101.3285$; $p = 0.9832$; $X_3^2 = 1.36$; $P = 0.72$;		$k = 175.1337$; $p = 0.9904$; $X_3^2 = 1.71$; $P = 0.64$.	

	Text 62 *		Text 64		Text 66 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	69	72.55	109	111.82	105	111.00
2	74	73.46	82	71.14	84	77.20
3	55	49.66	28	33.57	42	35.82
4	27	25.22	8	13.09	9	12.47
5	8	10.26	7	4.46	1	4.51
6	3	4.85	1	1.37	-	-
7	-	-	1	0.55	-	-
	$k = 640.2699$; $p = 0.9968$; $X_3^2 = 2.07$; $P = 0.56$.		$k = 7.8581$; $p = 0.8564$; $X_3^2 = 6.09$; $P = 0.11$;		$k = 1264.5662$; $p = 0.9989$; $X_2^2 = 5.66$; $P = 0.06$.	

	Text 67		Text 68 *		Text 69	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	117	122.04	121	122.16	108	109.40
2	71	69.91	72	65.57	72	66.27
3	33	29.19	17	24.26	25	29.61
4	12	9.92	9	6.95	10	10.87
5	-	2.91	2	2.06	5	3.47
6	1	0.76	-	-	1	1.38
7	-	0.18	-	-	-	-
8	-	0.04	-	-	-	-
9	1	0.05	-	-	-	-
	$k = 9.6914$; $p = 0.8928$; $X_2^2 = 2.09$; $P = 0.35$.		$k = 28.5402$; $p = 0.9637$; $X_2^2 = 3.42$; $P = 0.18$.		$k = 8.3954$; $p = 0.8711$; $X_3^2 = 2.06$; $P = 0.56$.	

Text 68: Obwohl der Aufsatz nach den Angaben des Verfassers nachträglich leicht verändert wurde, folgt er trotzdem dem vorgeschlagenen Modell. Es zeigt sich, daß Korrekturen nicht zwingend zur Verwerfung des Modells führen müssen.

	Text 70 *		Text 71		Text 72	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	87	87.54	113	113.24	72	69.82
2	53	52.11	67	66.44	48	49.35
3	21	20.95	27	26.96	33	30.35
4	6	6.39	8	8.50	16	17.27
5	2	2.01	3	2.86	5	9.35
6	-	-	-	-	6	4.89
7	-	-	-	-	4	2.49
8	-	-	-	-	2	2.48
	$k = 75.4706$; $p = 0.9844$; $X_2^2 = 0.04$; $P = 0.98$.		$k = 25.6655$; $p = 0.9560$; $X_2^2 = 0.04$; $P = 0.98$;		$k = 2.2761$; $p = 0.5685$; $X_5^2 = 3.69$; $P = 0.59$.	

	Text 73		Text 74		Text 75 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	129	130.88	74	74.00	106	108.03
2	77	65.24	42	43.60	73	68.30
3	15	26.57	24	21.45	26	29.11
4	10	9.61	9	9.51	10	9.41
5	2	3.21	2	3.94	-	2.46
6	2	1.01	3	1.55	3	0.69
7	2	0.48	1	0.95	-	-
	$k = 3.4271$; $p = 0.7748$; $X_2^2 = 7.58$; $P = 0.02$; $C = 0.0310$.		$k = 2.9583$; $p = 0.7023$; $X_2^2 = 0.42$; $P = 0.81$.		$k = 85.8599$; $p = 0.9854$; $X_2^2 = 0.74$; $P = 0.69$.	

Text 73: Dieser Aufsatz stimmt nur sehr schlecht mit dem Modell überein, denn es gilt $P = 0.02$ und $C = 0.0310$. Gründe für die schlechte Anpassung lassen sich nicht nennen.

	Text 76		Text 77 *		Text 78 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	49	47.60	84	86.26	61	64.10
2	49	52.36	84	75.25	63	55.86
3	42	40.21	40	46.43	36	33.58
4	28	24.21	20	22.72	8	15.65
5	10	12.16	13	9.38	6	6.03
6	3	5.30	4	4.96	3	2.00
7	3	2.06	-	-	1	0.79
8	-	0.72	-	-	-	-
9	1	0.38	-	-	-	-
	$k = 20.0732$; $p = 0.8956$; $X_5^2 = 2.75$; $P = 0.74$.		$k = 15.4064$; $p = 0.8937$; $X_3^2 = 3.87$; $P = 0.28$.		$k = 27.6054$; $p = 0.9391$; $X_3^2 = 5.51$; $P = 0.14$.	

Text 77: Das eingeschobene Gedicht „Tenebrae“ von Paul Celan wurde nicht mit ausgewertet, da bei Gedichten eine stärkere Konzentration auf die Form vorherrscht und die Satzlänge bewußt gestaltet ist. Eine Berücksichtigung des Gedichtes hätte dazu führen müssen, den Aufsatz als Mischtext anzusehen und entsprechende zusammengesetzte Verteilungen als Modell anzusetzen.

	Text 79		Text 80 *		Text 81	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	61	63.34	79	80.94	73	74.32
2	61	56.92	54	48.64	61	56.57
3	42	38.85	17	20.37	27	30.69
4	21	22.32	6	6.66	13	13.30
5	6	11.38	2	1.81	6	4.89
6	6	5.31	-	0.43	2	2.23
7	3	2.31	1	0.11	-	-
8	2	1.57	-	-	-	-
	$k = 6.1726$; $p = 0.7494$; $X_5^2 = 3.68$; $P = 0.60$.		$k = 21.8326$; $p = 0.9473$; $X_2^2 = 1.43$; $P = 0.49$.		$k = 13.4219$; $p = 0.8945$; $X_3^2 = 1.09$; $P = 0.78$.	

	Text 82 *		Text 83		Text 84 *	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	74	70.98	52	51.44	67	71.77
2	57	62.75	50	49.52	73	62.29
3	40	37.69	30	32.92	33	36.73
4	18	17.30	20	16.99	16	16.55
5	8	6.47	6	7.25	3	6.07
6	1	2.81	3	2.66	1	1.89
7	-	-	-	0.86	2	0.51
8	-	-	1	0.36	1	0.19
	$k = 51.1854$; $p = 0.9661$; $X_3^2 = 2.34$; $P = 0.50$.		$k = 26.7111$; $p = 0.9305$; $X_4^2 = 1.10$; $P = 0.89$.		$k = 51.1782$; $p = 0.9667$; $X_3^2 = 4.92$; $P = 0.18$.	

	Text 85	
x	n_x	NP_x
1	52	57.24
2	51	45.68
3	35	27.80
4	9	14.29
5	4	6.53
6	1	2.74
7	2	1.07
8	-	0.40
9	2	0.25
	$k = 5.9414$; $p = 0.7701$; $X_3^2 = 5.98$; $P = 0.11$.	

Einschränkend muß darauf hingewiesen werden, daß von den 78 Texten, für die eine Modellierung möglich war, einer den zulässigen Grenzwert von $C \leq 0.02$ überschreitet, und vier weitere Texte mit $P = 0.03$ bzw. $P = 0.02$ nur sehr kleine P-Werte aufweisen. Eine zusätzliche Betrachtung der C-Werte bei diesen Texten führt zu keiner Ergebnisverbesserung, denn in allen vier Fällen liegen diese Werte oberhalb der vorgegebenen Grenze, wenn auch bei Text 45 und 46 mit $C = 0.0285$ bzw. $C = 0.0278$ nur sehr knapp.

Für die Texte, die nicht dem Modell folgen, gilt, daß man mit einer Ausnahme anhand der empirischen Daten keine Abweichungen zu den Satzlängenverteilungen der übrigen Texte feststellen kann. Die Häufigkeitsverteilungen nehmen keinen Kurvenverlauf an, der es ermöglichen würde, aufgrund der beobachteten Werte darauf zu schließen, daß der Text dem Modell nicht folgt. Erst die theoretisch ermittelten Daten geben Aufschluß darüber, daß sich diese Texte nicht durch das vorgeschlagene Modell beschreiben lassen. Umgekehrt gilt aber auch, daß Texte, die in ihrer beobachteten Häufigkeitsverteilung geringfügig von dem erwarteten Häufigkeitsverlauf abweichen, indem sie einen etwa gleich hohen Anteil ein- und zweiclausiger Sätze aufweisen (z.B. Text 79), nicht zwangsläufig dem Modell nicht folgen. Die rechnerische Analyse zeigt, daß für diese Texte eine Anpassung an das Modell möglich ist, und man aufgrund der Auszählung noch keine Aussagen über die Güte der Anpassung machen kann.

Text 56 weicht als einziger Text mit seinem Auszählungsergebnis deutlich von den übrigen Texten ab. Die Vermutung, daß dieser Text nicht dem Modell der 0-gestutzten negativen Binomialverteilung folgt, bestätigt sich durch die rechnerische Analyse. „Das Grundrecht auf Kriegsdienstverweigerung" ist der einzige Aufsatz im Textkorpus, der von zwei Autoren erstellt wurde. Es ist anzunehmen, daß durch die Arbeit beider Autoren am Text dem Artikel keine einheitliche Struktur im Sinne unserer Annahme zugrunde liegt. Durch die Überlagerung der Stile beider Autoren wird der Text zu einem „Mischtext", was sich dadurch bestätigt, daß die Hyperpoissonverteilung

$$P_x = \frac{a^x}{{}_1F_1(1;b;a)b^{(x)}}, \quad x=0,1,\ldots; \quad a,b>0$$

ein geeignetes Modell für diesen Text darstellt.

Text 56

x	n_x	NP_x
1	36	37.06
2	64	65.88
3	55	47.03
4	16	21.00
5	6	6.82
6	1	1.74
7	1	0.36
8	1	0.11

$\alpha = 1.1927$;
$b = 0.6709$;
$X_2^2 = 3.03$;
$P = 0.39$.

Text 63 läßt sich ebenfalls weder durch die 0-gestutzte negative Binomialverteilung noch durch die positive Poissonverteilung anpassen. Die Gründe dafür liegen vermutlich darin, daß es sich bei dem Artikel um eine offizielle Stellungnahme des Instituts für Wirtschaftsprüfer handelt, die vermutlich mehrfach überarbeitet wurde, möglicherweise sogar von verschiedenen Personen. Auch hier erweist sich wiederum die Hyperpoissonverteilung als geeignetes Modell.

Text 63

x	n_x	NP_x
1	108	108.21
2	93	93.18
3	25	24.38
4	3	3.76
5	1	0.47

$\alpha = 0.3759$;
$b = 0.4366$;
$X_1^2 = 0.03$;
$P = 0.87$.

Ein Grund für die Diskrepanz zwischen Text 65 und dem Modell findet sich in einer Fußnote zum Text, die den Aufsatz als überarbeitete Fassung eines Vortrages ausweist. Durch die Überarbeitung ist der Artikel also nicht „in einem Zug" erzeugt. Es kommt zu Inhomogenitäten, und die theoretisch angenommene Verteilung erweist sich nicht als adäquates Modell für diesen Text[8]. Im Gegensatz zu den beiden vorangegangenen Texten, läßt sich dieser aber durch die positive Poissonverteilung darstellen, wenn $P = 0.03$ auch sehr klein ist. Für die Anpassung des Textes an die Hyperpoissonverteilung ergibt sich $P = 0.78$.

Text 65

x	n_x	NP_x
1	112	111.19
2	104	105.71
3	41	39.87
4	10	9.37
5	1	1.86

$\alpha = 0.6251$;
$b = 0.6251$;
$X_2^2 = 0.49$;
$P = 0.78$.

Die Häufigkeitsverteilungen der Satzlängen in den SPIEGEL-Artikeln 39, 42 und 43 folgen ebenfalls nicht dem Modell der 0-gestutzten negativen Binomialverteilung.

[8] Überarbeitete Texte führen aber nicht zwangläufig zur Verwerfung des Modells, denn Text 68 ist ebenfalls nachträglich leicht verändert worden, folgt in seiner Häufigkeitsverteilung aber dennoch der 0-gestutzten negativen Binomialverteilung.

lung. Alle drei Artikel weisen mehr oder weniger viele Zitate auf, bei denen anzunehmen ist, daß sie nicht original in deutscher Sprache gesprochen wurden, sondern vielmehr Übersetzungen aus anderen Sprachen darstellen[9]. Möglich ist, daß diese Übersetzungen fremdsprachiger Sätze zu Homogenitätsproblemen führen und das Modell somit nicht mit den beobachteten Daten übereinstimmt.

Ein sehr gutes Modell für Text 43 liefert die Conway-Maxwell-Poisson-Verteilung:

$$P_x = \frac{a^x}{(x!)^b T_1}, x = 1,2,\ldots, T_1 = \sum_{j=1}^{\infty} \frac{a^j}{(j!)^b}.$$

mit $P = 0.85$. Für Text 42 erhält man eine ausgezeichnete Anpassung mit Hilfe der Hyperpoissonverteilung, die jedoch leider keine Freiheitsgrade aufweist. Für den Text 39 ergibt sich eine Anpassung mit Hilfe der positiven Poissonverteilung.

	Text 39		Text 42		Text 43	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	109	115.29	108	107.98	69	69.29
2	67	54.53	75	74.98	58	57.94
3	15	17.16	8	8.48	14	14.00
4	-	4.06	-	0.52	2	1.77
5	1	0.93	-	0.02	-	-
6	-	-	1	0.02	-	-
	$\alpha = 0.9459$; $X_2^2 = 6.65$; $P = 0.04$.		$\alpha = 0.1351$; $b = 0.1946$; $C = 0.0000005$.		$\alpha = 0.8363$; $b = 1.7908$; $X_1^2 = 0.03$; $P = 0.85$.	

Für Text 35 ergibt ebenfalls eine zu große Abweichung vom Modell der 0-gestutzten negativen Binomialverteilung. Im Gegensatz zu den anderen sechs Texten kann hier jedoch keine naheliegende Begründung gegeben werden. Denkbar ist, daß besondere Textproduktionsbedingungen vorgelegen haben, die Inhomogenitäten hervorrufen und eine Anpassung an das Modell unmöglich machen. Überprüft man die Möglichkeit der Anpassung dieses Textes an die positive Poissonverteilung, so erhält man mit $P = 0.02$ zwar kein gutes, aber doch akzeptables Ergebnis:

[9] In Text 42 wird sogar explizit darauf hingewiesen, daß der Moskauer Atomphysiker kein Deutsch spricht („Obwohl er kein Deutsch spricht...", SPIEGEL, Nr. 16, S.222).

Text 35

x	n_x	NP_x
1	111	115.46
2	58	45.67
3	5	12.04
4	1	2.38
5	1	0.45
	$\alpha = 0.7911$; $X_2^2 = 7.86$; $P = 0.02$.	

Als Fazit der Untersuchung kann festgehalten werden, daß aufgrund der erzielten Resultate die negative Binomialverteilung in ihrer 0-gestutzten Form als Modell der Satzlängenhäufigkeitsverteilung, gemessen in der Anzahl der Teilsätze, bis auf weiteres beibehalten werden kann, da nur wenige Texte nicht der Verteilung folgen und sich in der Regel Gründe für die Abweichungen nennen lassen. Bei den mit * gekennzeichneten Texten erweist sich zusätzlich auch die positive Poissonverteilung als gutes Modell der Satzlängenverteilung und liefert dann zumeist etwas höhrere P-Werte als die 0-gestutzte negative Binomialverteilung. Da die positive Poissonverteilung jedoch nur einen Grenzfall der 0-gestutzten negativen Binomialverteilung für sehr großes k und p fast 1 darstellt und nur für solche Parameter als Modell in Frage kommt, bleibt die 0-gestutzte negative Binomialverteilung für den allgemeinen Fall das bessere Modell.

Die Tatsache, daß sich nicht alle Texte durch dieses Modell beschreiben lassen und die darauf resultierende Notwendigkeit der Anpassung von vier verschiedenen Funktionen läßt sich dahingehend interpretieren, daß jeder Text an gewisse Bedingungen gebunden ist, die die Form der Texte bestimmen und sie modifizieren können. Aufgrund verschiedener Rand- und Ausgangsbedingungen der einzelnen Texte können unterschiedliche Modellverteilungen nötig werden, die aber als verschiedenartige Ausprägungen ein und derselben Gesetzmäßigkeit anzusehen sind (vgl. Wimmer, Köhler, Grotjahn & Altmann, 1994).

Auch sind Abweichungen vom Modell nicht nur negativ zu beurteilen, denn sie führen zu neuen Erkenntnissen und treiben die Theorie voran. „Every discrepancy between a law formula and empirical findings, if interpreted in the light of some hypotheses, becomes a new source of information rather than being just unfavorable or negative evidence" (Bunge, 1967:347).

Anhand dieses Ergebnisses wird deutlich, daß die Satzlänge kein chaotisches Merkmal ist und die Produktion von Texten keineswegs so willkürlich verläuft, wie man gemeinhin annimmt. Die Bestätigung der Existenz theoretischer Modelle, vor allem der 0-gestutzten negativen Binomialverteilung, zeigt, daß nicht nur

grammatische Regeln gelten, sondern auch stochastische Gesetzmäßigkeiten auf den Text wirken, denen der Texterzeuger bei der Produktion unbewußt folgt.

5. Die Beziehung der Parameter *q* und *k*

Ein zweites wichtiges Ergebnis der Untersuchung ergibt sich aus der Betrachtung der Parameter *q* und *k*. Bereits ihre Interpretation

$$k - 1 := \frac{a}{b} = \frac{Text}{Sprecher}, \qquad q := \frac{b}{c} = \frac{Sprecher}{Hörer}$$

legt die Vermutung nahe, daß die Wahl nicht zufällig erfolgt. Die Parameter erweisen sich als voneinander abhängig, denn durch seine Beziehung zum Autor kann der Hörer beispielsweise Einfluß auf die Gestaltung des Textes nehmen, indem er die Möglichkeiten des Autors durch seine Forderungen einschränkt (vgl. Altmann & Best, 1996). Faßt man *k* als Funktion von *q* auf, d.h. $k = f(q)$, und werden die Zahlenpaare (q, k) in ein karthesisches Koordinatensystem eingetragen, so bestätigt sich diese Vermutung der Abhängigkeit zwischen *q* und *k*, denn man erhält als Trend eine fallende Kurve, die für kleines *q* sehr große Werte annimmt und umgekehrt für großes *q* kleine Werte aufweist.

Sieht man die Hörergemeinschaft als die Instanz an, die dem Sprecher übergeordnet ist (vgl. Altmann & Best, 1996), so ist zu überprüfen, ob die Parameterbeziehung möglicherweise der gleichen Gesetzmäßigkeit folgt, die auch das Menzerathsche Gesetz steuert. Geht man von der Formel $\hat{k} = cq^d$ aus, so lassen sich mit Hilfe der Daten die Parameter *c* und *d* iterativ berechnen, und man erhält dann in diesem Fall die Gleichung $\hat{k} = 0.4126q^{-1.647}$.

Tabelle 6 enthält die Zuordnung der empirisch ermittelten *k* sowie der theoretisch berechneten $\hat{k}$ zu den entsprechenden *q*'s. Zur besseren Übersicht sind die *q*'s der Größe nach geordnet und *k* und $\hat{k}$ auf zwei Stellen gerundet.

Vergleicht man die Ergebnisse in der Tabelle, so zeigt sich, daß die Daten mit sehr großem *q* und extrem kleinen *k* fast alle zu literarischen Texten, vornehmlich aus dem Bereich der Kurzprosa für Erwachsene, gehören. Für alle übrigen Textgruppen läßt sich ein solcher Zusammenhang mit der Größe der Parameter jedoch nicht erkennen.

Tabelle: Zuordnung von k und $\hat{k}$ zu q

Text	$q = 1 - p$	k	$\hat{k} = cq^d$
44	0.0010	1110.38	1287.05
66	0.0011	1264.57	1151.83
57	0.0012	1233.22	1040.82
50	0.0015	704.32	802.61
45	0.0016	731.45	744.49
40	0.0017	446.88	693.73
55	0.0017	833.67	693.73
53	0.0019	714.87	609.44
59	0.0019	564.41	609.44
47	0.0021	322.50	542.38
21	0.0026	526.10	422.94
14	0.0030	376.52	358.01
10	0.0032	206.90	332.08
62	0.0032	640.27	332.08
46	0.0036	329.83	289.51
9	0.0043	148.37	235.39
5	0.0057	177.40	169.52
49	0.0061	287.62	156.65
41	0.0066	177.90	142.91
6	0.0074	138.65	125.08
17	0.0079	143.76	115.91
18	0.0080	142.09	114.23
61	0.0096	175.13	92.37
4	0.0107	94.60	81.41
36	0.0107	133.60	81.41
3	0.0127	88.78	66.68
75	0.0146	85.86	56.69
70	0.0156	75.47	52.48
60	0.0168	101.33	48.14
19	0.0206	55.97	37.96
13	0.0292	19.27	25.28
84	0.0333	51.18	21.70
82	0.0339	51.19	21.25
68	0.0363	28.54	19.62
26	0.0397	28.23	17.68
51	0.0436	25.19	15.85
71	0.0440	25.67	15.68
31	0.0466	24.97	14.67
2	0.0497	17.82	13.61
1	0.0511	8.86	13.18
80	0.0527	21.83	12.71
78	0.0609	27.61	10.74

Text	$q = 1 - p$	k	$\hat{k} = cq^d$
83	0.0695	26.71	9.21
7	0.0973	4.07	6.22
8	0.0992	7.35	6.08
54	0.1029	14.92	5.83
76	0.1044	20.07	5.73
81	0.1055	13.42	5.66
77	0.1063	15.41	5.61
67	0.1072	9.69	5.56
15	0.1211	7.18	4.82
16	0.1275	4.61	4.54
69	0.1289	8.40	4.49
25	0.1333	8.35	4.31
64	0.1436	7.86	3.96
52	0.1585	5.94	3.53
38	0.1739	5.01	3.16
37	0.1822	6.12	3.00
73	0.2252	3.43	2.34
48	0.2275	1.58	2.31
85	0.2299	5.94	2.29
58	0.2405	1.55	2.17
79	0.2506	6.17	2.07
29	0.2676	0.74	1.92
52 a	0.2743	2.32	1.86
74	0.2972	2.96	1.69
24	0.3333	3.03	1.48
22	0.3501	3.12	1.40
72	0.4315	2.28	1.10
20	0.4379	2.20	1.08
27	0.4448	0.05	1.06
11	0.4971	0.51	0.93
28	0.4975	0.02	0.93
23	0.5036	1.37	0.92
34	0.5199	1.35	0.88
12	0.5280	0.92	0.87
30	0.5422	0.73	0.84
32	0.7003	0.01	0.62
33	0.7213	0.76	0.60

Da die Werte für k und $\hat{k}$ sehr großen Schwankungen unterliegen, d.h. für $q \to 0$ extrem groß und für $q \to 1$ sehr klein werden, wird zur besseren Übersicht die Kurvendarstellung aufgesplittet. Abbildung 1a enthält den empirisch ermittelten

und den theoretisch berechneten Kurvenverlauf für $q < 0.02$, und die Abbildung 1b stellt die Kurven für $q > 0.02$ dar.

Aus der Tabelle 1 und Abbildung 1a ergibt sich, daß für kleines q die praktisch beobachteten Werte nur verhältnismäßig schlecht mit den theoretisch ermittelten übereinstimmen und sich so beider Kurven nur bruchstückhaft gleichen. Dies hängt vermutlich damit zusammen, daß für sehr kleines q und großes k die Verteilung der 0-gestutzten negativen Binomialverteilung durch die positive Poissonverteilung approximiert werden kann. Für großes q zeigt Abbildung 1b dagegen nur geringe Abweichungen der beiden Kurven, so daß sich insgesamt eine recht gute Übereinstimmung zu ergeben scheint.

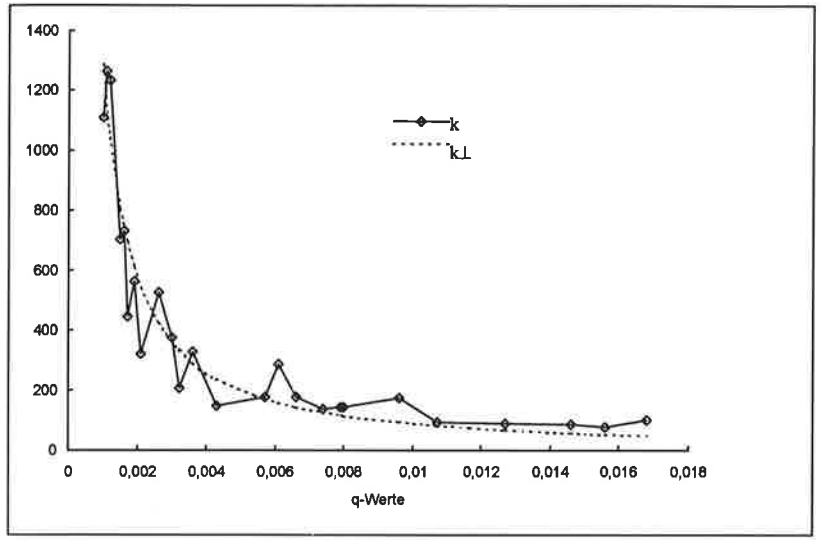


Abb. 1a: Darstellung der Kurven k und $\hat{k}$ für $q < 0.02$

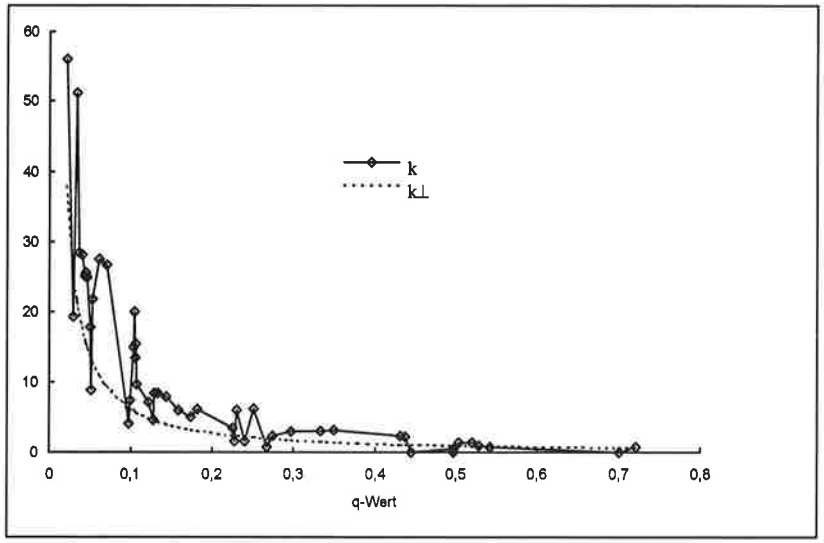


Abb. 1b: Darstellung der Kurven k und $\hat{k}$ für $q > 0.02$

Der optisch gewonnene Eindruck der insgesamt guten Übereinstimmung der Kurven läßt sich mit Hilfe des Determinationskoeffizienten D, durch den die Güte der Anpassung rechnerisch bestimmt werden kann, bestätigen. Es gilt

$$D = 1 - \frac{\sum_{i=1}^{n}\left(k_i - \hat{k}_i\right)^2}{\sum_{i=1}^{n}\left(k_i - \tilde{k}\right)^2}$$

mit $\tilde{k}$ = *Durchschnitt aller* k_i und n = *Anzahl der Texte*. Das Resultat wird als gut angesehen, wenn $D > 0.9$ gilt. Da für sieben Texte keine Anpassungen an das Modell vorgenommen werden konnten und ein Text doppelt ausgewertet wurde, ergibt sich $\tilde{k}$ = 159.10 und n = 78. Als Determinationskoeffizient erhält man damit D = 0.9384, was ein sehr gutes Ergebnis darstellt. Bedenkt man, wie groß die Abweichungen zwischen k und $\hat{k}$ für kleines q sind, so läßt sich das Ergebnis noch verbessern, wenn diejenigen Zahlenpaare (q, k) herausgenommen werden, deren Texte sich durch die positive Poissonverteilung anpassen lassen. Es ergibt sich dann D = 0.9573.

Insgesamt kann festgehalten werden, daß die Wahl der Parameter nicht willkürlich zu erfolgen scheint, sondern hier vermutlich der gleiche Zusammenhang

wirksam ist, der auch das Menzerathsche Gesetz steuert. Um das Ergebnis zu untermauern, müssen weitere Untersuchungen durchgeführt werden.

6. Ausblick

Die Untersuchung läßt sich auf verschiedene Weise fortführen oder mit dem Experiment zur Wortlänge verknüpfen. Von den vorhandenen Möglichkeiten seien hier einige erwähnt.

Für mittellange deutsche Texte bestätigte sich die Hypothese, daß die 0-gestutzte negative Binomialverteilung ein geeignetes Modell der Satzlängenverteilung ist. Das Experiment zur Wortlänge, das derzeit in Göttingen durchgeführt wird, versucht, für deutsche Texte dieselbe theoretische Verteilung als Modell der Wortlängen zu bekräftigen. Beide Untersuchungen messen die Länge ihrer Konstrukte durch die Anzahl der Elemente der nächst niedrigeren Ebene. Die verallgemeinerte Annahme, daß die Verteilung der Längen eines Konstrukts, gemessen durch die Anzahl seiner Komponenten, dem Modell der 0-gestutzten negativen Binomialverteilung folgt, müßte durch weitere Untersuchungen von Konstrukt-Komponenten-Beziehungen wie z.B. der Clause-Wort-Verteilung gefestigt werden.

Eine andere Möglichkeit, die Untersuchung fortzusetzen, besteht in der Erweiterung des Textkorpus auf Texte aus anderen Jahrhunderten. Bei der Auswertung der Auszählungsergebnisse wurde mehrfach auf das Vordringen des Nominalstils verbunden mit der Auflösung hypotaktischer Satzstrukturen hingewiesen. Im 19. Jahrhundert herrschte dagegen der Verbalstil vor, so daß durchaus denkbar ist, daß ältere Texte eine andere Häufigkeitsverteilung aufweisen, die aber möglicherweise lediglich verschoben ist. Mit der Überlegung der möglicherweise veränderten Verteilung einher geht die Frage nach der Akzeptanz der 0-gestutzten negativen Binomialverteilung als Modell nichtgegenwartssprachlicher Texte.

Die vorliegende Untersuchung beschränkte sich auf Texte mittlerer Länge, da davon ausgegangen werden muß, daß Textgesetze nicht für beliebig lange Texte gelten. Eine Abgrenzung nach oben erscheint sinnvoll, wenn man bedenkt, daß es fast unmöglich ist, sehr lange Texte „in einem Zug" zu erzeugen, gleichzeitig aber stellt sich die Frage, ob eine Beschränkung nach unten nötig ist. Eine Ausdehnung des Textkorpus auf sehr kurze Texte wie Zeitungsartikel oder Behördentexte könnte Aufschluß darüber geben, ob auch solche Texte dem Modell folgen und die Voraussetzung einer Mindestgrenze somit überflüssig ist. Interessant wäre in diesem Zusammenhang auch die Untersuchung der Satzlängen bei Gedichten. Neben der in der Regel geringen Anzahl an Sätzen kommt für Gedichte eine stärkere Konzentration des Verfassers auf formale Merkmale hinzu. Durch die Anpassung an den Rhythmus und das Metrum erscheint es denkbar, daß hier eine bewußte Gestaltung der Satzlänge erfolgt. Um spezielle Effekte zu erwirken,

unterwirft sich der Autor möglicherweise anderen Gesetzen, die dem Modell der 0-gestutzten negativen Binomialverteilung entgegenwirken. Altmann überprüfte 1988 in seinem Aufsatz die negative Binomialverteilung als Modell der Satzlänge bereits an Texten verschiedener Sprachen. Da insgesamt nur zehn Texte sieben verschiedener Sprachen untersucht wurden, ist es nötig, das Ergebnis durch weitere Forschungsarbeiten zu festigen. Eine Erweiterung auf Sprachen, für die die Untersuchung zur Wortlänge gezeigt hat, daß die Entwicklung anderer Modelle nötig wird, wenn die Wortlängenhäufigkeit durch Sprachrhythmus oder ähnliches überlagert wird (vgl. Best & Zhu,1994:28), könnte aufzeigen, ob auch die Satzlänge von Faktoren wie dem Sprachrhythmus abhängt und in solchen Fällen sich wiederum die 0-gestutzte negative Binomialverteilung nicht als Modell bestätigt. Gleichzeitig ließe ein Vergleich der Wortlänge und der Satzlänge möglicherweise den Rückschluß zu, daß beide Verteilungen in einer Sprache immer demselben Modell folgen.

Ein letzter Punkt, der hier angesprochen werden soll, bezieht sich auf die Möglichkeit der unterschiedlichen Wahl der Einheiten, mittels derer die Satzlänge bestimmt werden kann. Während es zunächst sinnvoll erscheint, den Aufbau einer gegebenen Einheit durch Einheiten der nächst niedrigeren Ebene zu beschreiben, um Störfaktoren möglichst gering zu halten, bietet sich aber auch immer die Möglichkeit, größere hierarchische Abstände zu wählen und sehr kleine Einheiten zur Beschreibung zu verwenden. Für die Satz-Wort-Verteilung existieren bereits Modelle, dagegen sind bislang beispielsweise für die Messung in Phonemen oder Silben noch keine theoretischen Verteilungen entwickelt worden. Ehe jedoch diese Arbeit aufgenommen werden kann, ist es sinnvoll, zunächst zu klären, ob solche Messungen überhaupt noch brauchbare Ergebnisse liefern oder ob nicht vielmehr die Störfaktoren, die sich aus allen Störungen der zwischen gelegenen Ebenen zusammensetzen, bereits so groß sind, daß das Ergebnis zu stark beeinträchtigt wird, um damit noch sinnvolle Aussagen machen zu können. Liefern die Messungen noch brauchbare Ergebnisse, so wäre im nächsten Schritt zu fragen, ob der Ansatz von Altmann, den er sowohl für das Modell der negativen Binomialverteilung als auch für die Hyperpascalverteilung verwendet hat, sich so modifizieren läßt, daß er sich auch für die Herleitung anderer Modelle als sinnvoll erweist, oder ob ein ganz neuer Ansatz gefunden werden muß, um passende theoretische Verteilungen zu entwickeln (vgl. auch Altmann, 1988:156). Möglich ist auf jeden Fall die weitere Überprüfung der von Altmann vorgeschlagenen Hyperpascalverteilung als Modell der Satz-Wort-Variante. Eine erneute Auszählung der bereits zur Bestätigung der 0-gestutzten negativen Binomialverteilung verwendeten 85 Texte würde gleichzeitig die Möglichkeit zum Vergleich bieten. Es ließe sich vielleicht die Frage klären, ob Texte, für die das eine Modell akzeptiert werden kann, auch immer dem anderen Modell folgen und somit ein Zusammenhang zwischen der negativen Binomialverteilung als Modell der Satz-Clause-Zählung und der Hyperpascalverteilung als Modell der Satz-Wort-

Zählung erkennbar ist. Denkbar ist in diesem Zusammenhang auch die Kombination der beiden Meßeinheiten. Man könnte die Satzlängenverteilung dann gleichzeitig durch die Anzahl der Teilsätze *und* die Anzahl der Wörter darstellen und versuchen, für solche Verteilungen passende mehrdimensionale Modelle zu entwickeln.

Quellentexte

1. Kurzprosa für Kinder:

Borchers, E.: Reise mit Samuel. In: Pregel, D. u.a. (Hrsg.): *Texte für die Primarstufe. 3. Schuljahr*. Hannover 1973, 123-129.

Bröger, A.: Moritz tauscht. In: Bröger, A.: *Moritzgeschichten*. Ravensburg 1983, 72-78.

Bröger, A.: Moritz lernt das Schwimmen. In: Bröger, A.: *Moritzgeschichten*. Ravensburg 1983, 6-11.

Härtling, P.: Der Ausreißer. In: Härtling, P.: *Geschichten für Kinder*. Weinheim u.a. 1988, 9-19.

Kordon, K.: Zahl oder Adler. In: *Die schönsten Freundschaftsgeschichten* (hrsg. von H. Westhoff). Ravensburg 1987, 228-235.

Krenzer, R.: Die Sache mit der Schultasche. In: *Meine Welt. Geschichten zum Lesen und Vorlesen* (hrsg. von I. Ryssel). Gütersloh 1988, 12-17.

Lengauer, M.: Geschichte einer Großmutter. In: *Oma-Geschichten*. Wien 1986, 62-70.

Pausewang, G.: Sascha und Elisabeth. In: *Die schönsten Freundschaftsgeschichten* (hrsg. von H. Westhoff). Ravensburg 1987, 11-18.

Pelz, M.: Der Wind in der Krottenbachstraße. In: *Meine Welt. Geschichten zum Lesen und Vorlesen* (hrsg. von I. Ryssel). Gütersloh 1988, 71-79.

Prokop, G.: Die Maus im Fenster. In: Prokop, G.: *Die Maus im Fenster. Gute-Nacht-Geschichten*. Zürich, Köln 1982, 7-18.

Rettich, M.: Geschichte ohne Ende. In: Rettich, M.: *Allerlei von früher, jetzt und irgendwo*. Hamburg 1986, 97-105.

Rettich, M.: Die Landstraßengeschichte. In: Rettich, M.: *Allerlei von früher, jetzt und irgendwo*. Hamburg 1986, 115-121.

Rettich, M.: Michels Kaninchen. In: Rettich, M.: *Seidenhund und Lumpenköter. Geschichten von besonderen Tieren*. Wien, München 1988, 22-30.

Rettich, M.: Trines Spuk. In: Rettich, M.: *Seidenhund und Lumpenköter. Geschichten von besonderen Tieren*. Wien, München 1988, 126-132.

Schultheis, O.: Zirkus auf dem Bauernhof. In: *Ich hör' so gern Geschichten: kleine Geschichten zum Vorlesen* (hrsg. von U. Schultheiss). München 1988, 37-46.
Wöffel, U.: Das Miststück. In: Wölfel, U.: *Die grauen und die grünen Felder. Wahre Geschichten*. Mühlhein/Ruhr 1970, 79-86.
Wöffel, U.: In einem solchen Land. In: Wölfel, U.: *Die grauen und die grünen Felder. Wahre Geschichten*. Mülheim/Ruhr 1970, 51-57.

2. Kurzprosa für Erwachsene:

Aichinger, I.: Die geöffnete Order. In: Aichinger, I.: *Meine Sprache und ich. Erzählungen.* Frankfurt/M. 1981, 20-26.
Aichinger, I.: Engel in der Nacht. In: Aichinger, I.: *Meine Sprache und ich. Erzählungen.* Frankfurt/M. 1981, 38-45.
Bender, H.: Die Wölfe kommen zurück. In: *Erzählte Zeit. 50 deutsche Kurzgeschichten der Gegenwart* (hrsg. von M. Durzak). Stuttgart: Reclam 1980, 183-189.
Bender, H.: In der Gondel. In: Bender, H.: *Mit dem Postschiff. 24 Geschichten.* München 1962, 10-14.
Bender, H.: Mein Onkel aus Amerika. In: Bender, H.: *Der Hund von Torcello. 32 Geschichten*. Frankfurt 1984, 164-168.
Böll, H.: Der Mann mit den Messern. In: Böll, H.: *Wanderer kommst du nach Spa... Erzählungen*. 29. Aufl., München 1987, 1-26.
Elsner, G.: Die Mieterhöhung. In: *Erzählungen seit 1960 aus der Bundesrepublik Deutschland, aus und der Schweiz* (hrsg. von H. Vormweg). Stuttgart: Reclam 1983, 206-215.
Gaiser, G.: Der Schlangenkönig. In: Gaiser, G.: *Mittagsgesicht. Erzählungen*. Ostfildern 1983, 125-132.
Lenz, S.: Die Flut ist pünktlich. In: Lenz, S.: *Die Erzählungen (1949-1984).* Bd. 1 (1949-1958). München 1986, 65-70.
Lenz, S.: Ein Haus aus lauter Liebe. In: Lenz, S.: *Die Erzählungen (1949-1984).* Bd. 1 (1949-1958). München 1986, 57-64.
Reinig, C.: Die Wölfin. In: Reinig, C.: *Gesammelte Erzählungen*. Darmstadt u. Neuwied 1986 (Sammlung Luchterhand), 288-293.
Rinser, L.: Die rote Katze. In: *Erzählte Zeit. 50 deutsche Kurzgeschichten der Gegenwart* (hrsg. von M. Durzak). Stuttgart: Reclam 1980, 83-90.
Schnurre, W.: Auf der Flucht. In: *Erzählte Zeit. 50 deutsche Kurzgeschichten der Gegenwart (hrsg. von M. Durzak).* Stuttgart: Reclam 1980,199-203.
Walser, M.: Der Umzug. In: Walser, M.: *Gesammelte Geschichten*. Frankfurt/M. 1983, 22-32.

Walser, M.: Gefahrenvoller Aufenthalt. In: Walser, M.: *Ein Flugzeug über dem Haus und andere Geschichten.* Frankfurt/M. 1963, 14-25.
Widmer, U.: Tod und Sehnsucht. In: *Erzählungen seit 1960 aus der Bundesrepublik Deutschland, aus Österreich und aus der Schweiz* (hrsg. von H. Vormweg). Stuttgart: Reclam 1983, 194-201.
Wohmann, G.: Wiedersehen in Venedig. In: Wohmann, G.: *Ausgewählte Erzählungen aus zwanzig Jahren.* Bd. 1. Darmstadt und Neuwied 1979 (Sammlung Luchterhand), 25-33.

3. Artikel aus dem Nachrichtenmagazin DER SPIEGEL:

Da nicht immer ein Verfasser angegeben ist, werden die SPIEGEL-Artikel nach Sparten aufgelistet und jede Sparte chronologisch geordnet.

SPARTE WIRTSCHAFT:
Service nur für Steuersünder. DER SPIEGEL, Nr. 22, 47. Jahrgang, 1993, 114-119.
SPARTE DEUTSCHLAND:
Doler, I.: „Klagt nicht an der Mauer". DER SPIEGEL, Nr. 29, 47. Jahrgang, 1993, 52-61.
SPARTE KULTUR:
Enzensberger, H.M.: Ausblicke auf den Bürgerkrieg. DER SPIEGEL, Nr. 25, 47. Jahrgang, 1993, 170-175.
SPARTE PROZESSE:
Friedrichsen, G.: Hört das nie auf? DER SPIEGEL, Nr. 26, 47. Jahrgang, 1993, 85-89.
SPARTE AUSLAND:
Preuss, J.: Der gedemütigte Held. DER SPIEGEL, Nr. 16, 47. Jahrgang, 1993, 218-224.
Mayr, W.: „Sie sind zäh wie die Teufel". DER SPIEGEL, Nr. 22, 47. Jahrgang, 1993, 160-166.
Mayr, W.: Im Zahnrad der Zeit. DER SPIEGEL, Nr. 25, 47. Jahrgang, 1993, 122-126.
Mayr, W.: Totentanz im Ghetto. DER SPIEGEL, Nr. 28, 47. Jahrgang, 1993, 117-120.
Waffen für den Todfeind. DER SPIEGEL, Nr. 45, 47. Jahrgang, 1993, 199-208.
SPARTE TITEL:
Recycling ist nur der zweitbeste Weg. DER SPIEGEL, Nr. 25, 47. Jahrgang, 1993, 34-48.
X für unbekannt. DER SPIEGEL, Nr. 25, 47. Jahrgang, 1993, 160-169.
Die Bank gewinnt immer. DER SPIEGEL, Nr. 34, 47. Jahrgang, 1993, 90-97.

Kämpfen und Kungeln. DER SPIEGEL, Nr. 43, 47. Jahrgang, 1993, 50-63.
Augstein, Rudolf: „Freunde für immer". DER SPIEGEL, Nr. 1, 48. Jahrgang, 1994, 84-94.
SPIEGEL-Spezial:
Dahrendorf, R.: Eine große, universelle Sicht. SPIEGEL-Spezial, Nr. 4, 1993, 7-12.
Dörler, B.: „Unser Marsch hat begonnen". SPIEGEL-Spezial, Nr. 4, 1993, 34-40.
Lepenies, W.: Vorwärts mit der Aufklärung. SPIEGEL-Spezial, Nr. 4, 1993, 88-96.

4. Fachwissenschaftliche Texte:

Rechtswissenschaftliche Texte:

Dörr, D.: Der „numerus clausus" und die Kapazitätskontrolle durch die Verwaltungsgerichte. *Juristische Schulung. Zeitschrift für Studium und praktische Ausbildung*. 28. Jahrgang, Heft 2/1988, 96-102.
Friedrichs, H.-J.: Das neue Betreuungsgesetz. *Monatsschrift für Deutsches Recht*. 46. Jahrgang, Heft 1/1992, 1-9.
Lorenz, F.L.: Beschlagnahme von Krankenunterlagen - Prozessuale Anmerkungen zur Memmingen-Entscheidung des BGH. *Monatsschrift für Deutsches Recht*. 46. Jahrgang, Heft 4/1992, 313-318.
Peter, C. & Ludwig, R.: Das Grundrecht auf Kriegsdienstverweigerung. *Monatsschrift für Deutsches Recht*. 45. Jahrgang, Heft 12/1991, 1105-1110.
Ruhwedel, E.: Grundlagen und Rechtswirkungen sogenannter relativer Verfügungsverbote. *Juristische Schulung. Zeitschrift für Studium und praktische Ausbildung*. 20. Jahrgang, Heft 3/1980, 161-168.
Strauch, D.: Rechtsgrundlagen der Haftung für Rat, Auskunft und Gutachten. *Juristische Schulung. Zeitschrift für Studium und praktische Ausbildung*. 32. Jahrgang, Heft 11/1992, 897-902.

Wirtschaftswissenschaftliche Texte:

Hofmann, R.: Zusammenarbeit zwischen Interner Revision und Abschlußprüfung - Instrumente zur Erhöhung des Wirkungsgrades. *Der Betrieb. Wochenschrift für Betriebswirtschaft, Steuerrecht, Wirtschaftsrecht, Arbeitsrecht*. 44. Jahrgang, Heft 44/1991, 2249-2254.
Jebens, C.T.: Rückstellungen aus schwebende Waren-Einkaufskontrakte für einen unterdurchschnittlichen Unternehmergewinn. *Der Betrieb*... 42. Jahrgang, Heft 3/1989, 133-136.

Lachnit, L.: Erfolgs- und Finanzplanung für mittelständische Betriebe als Electronic-Banking-Leistung der Kreditinstitute. *Der Betrieb* ... 44. Jahrgang, Heft 42/1991, 2145-2152.
Paschek, W.: Chancen und Forderungen für die rückgedeckte Gruppenunterstützungskasse. *Der Betrieb*... 44. Jahrgang, Heft 17/1991, 873-878.
Wagner, F.W.: Perspektiven der Steuerberatung: Steuerrechtspflege oder Planung der Steuervermeidung? *Der Betrieb*... 44. Jahrgang, Heft 1/1991, 1-7.
Vahl, G.: Die Stellungnahme des Instituts der Wirtschaftsprüfer zur Unternehmensbewertung. *Der Betrieb*... 37. Jahrgang, Heft 23/1984, 1205-1209.

Geschichtswissenschaftliche Texte:

Bookmann, H.: Die Vergangenheit des Deutschen Ordens im Dienste der Gegenwart. *Geschichte in Wissenschaft und Unterricht. Zeitschrift des Verbandes der Geschichtslehrer Deutschlands*. 41. Jahrgang, Heft 6/1990, 370-385.
Erdmann, K.D.: Das Grundgesetz in der Verfassungsgeschichte. *Geschichte in Wissenschaft und Unterricht*... 30. Jahrgang, Heft 12/1979, 720-732.
Jordan, K.: Die Gestalt Heinrichs des Löwen im Wandel des Geschichtsbildes. *Geschichte in Wissenschaft und Unterricht*... 26. Jahrgang, Heft 4/1975, 226-241.
Röper, E.: Die Verfassung des Deutschen Bundes. *Geschichte in Wissenschaft und Unterricht*... 28. Jahrgang, Heft 11/1977, 648-668.
Steinbach, P.: Perspektiven politischer Integration. *Geschichte in Wissenschaft und Unterricht*... 35. Jahrgang, Heft 7/1984, 434-452.

5. Philosophische Texte:

Gadamer, H.-G.: Apologie der Heilkunst. In: Gadamer, H.-G.: *Gesammelte Werke*, Bd. 4: Neuere Philosophie II. Probleme und Gestalten. Tübingen 1987, 267-275.
Gadamer, H.-G.: Der Tod als Frage. In: Gadamer, H.-G.: *Gesammelte Werke*, Bd. 4: Neuere Philosophie II. Probleme und Gestalten. Tübingen 1987, 161-172.
Gadamer, H.-G.: Philosophische Bemerkungen zum Problem der Intelligenz. In: Gadamer, H.-G.: *Gesammelte Werke*, Bd. 4: Neuere Philosophie II. Probleme und Gestalten. Tübingen 1987, 276-287.
Gehlen, A.: Die gewaltlose Lenkung. In: Gehlen, A.: *Gesamtausgabe*, Bd. 7: Einblicke (hrsg. von Karl-Siegbert Rehberg). Frankfurt/M. 1978, 296-310.
Horkheimer, M.: Der Mensch in der Wandlung seit der Jahrhundertwende. In: Horkheimer, M.: *Gesammelte Schriften* (hrsg. von A. Schmidt, G. Schmid Noerr), Bd. 8: Vorträge und Aufzeichnungen 1949-1973. Frankfurt/M. 1985, 131-142.

Horkheimer, M.: Soziologie und Philosophie. In: Horkheimer, M.: *Gesammelte Schriften* (hrsg. von A. Schmidt, G. Schmid Noerr), Bd. 7: Vorträge und Aufzeichnungen 1949-1973. Frankfurt/M. 1985, 108-121.

Horkheimer, M.: Theismus und Atheismus. In: Horkheimer, M.: *Gesammelte Schriften* (hrsg. von A. Schmidt, G. Schmid Noerr), Bd. 7. Vorträge und Aufzeichnungen 1949-1973. Frankfurt/M. 1985, 173-186.

Jaspers, K.: Die Aufgabe der Philosophie in der Gegenwart. In: Jaspers, K.: *Was ist Philosophie?* München 1976, 129-137.

Jaspers, K.: Die Idee des Artzes. In: Jaspers, K.: *Philosophie und Welt. Reden und Aufsätze.* München 1958, 169-183.

Jaspers, K.: Die Philosophie in der Welt. In: Jaspers, K.: *Was ist Philosophie?* München 1976, 121-128.

Plessner, H.: Das Lächeln. In: Plessner, H.: *Gesammelte Schriften* (hrsg. von G. Dux u.a.), Bd. 7: Ausdruck und menschliche Natur. Frankfurt/M. 1982, 419-434.

Plessner, H.: Die Musikalität der Sinne. Zur Geschichte eines modernen Phänomens. In: Plessner, H.: *Gesammelte Schriften* (hrsg. von G. Dux u.a.), Bd. 7: Ausdruck und menschliche Natur. Frankfurt/M. 1982, 479-492.

Plessner, H.: Der Weg der Soziologie in Deutschland. In: Plessner, H.: *Gesammelte Schriften* (hrsg., von G. Dux u.a.), Bd. 10: Schriften zur Soziologie und Sozialphilosophie. Frankfurt/M.1985, 191-211.

Popper, K.: Selbstbefreiung durch das Wissen. In: Popper, K.: *Auf der Suche nach einer besseren Welt. Vorträge und Aufsätze aus dreißig Jahren.* München 1984, 149-163.

Weizsäcker von, C. F.: Die Verteidigung der Freiheit. In: Weizsäcker von, C.F.: *Deutlichkeit. Beiträge zur politischen und religiösen Gegenwart*. München u.a. 1978. 9-21.

Weizsäcker von, C. F.: Kirchenlehre und Weltverständnis. In: Weizsäcker von, C.F.: *Deutlichkeit. Beiträge zur politischen und religiösen Gegenwart*. München u.a. 1978,137-153.

Weizsäcker von, C.F.: Wozu Meditation? In: Weizsäcker von, C. F..: *Wahrnehmung der Neuzeit.* München u.a. 1983 (3. Aufl.), 316-324.

Literatur

Altmann, G. (1972). Status und Ziele der quantitativen Sprachwissenschaft. In S. Jäger (Hg.), *Linguistik und Statistik* (S. 1-9), Braunschweig: Vieweg.

Altmann, G. (1978). Towards a theory of language. In G. Altmann (Hg.), *Glottometrika 1* (S. 1-25), Bochum: Brockmeyer.

Altmann, G. (1983). H. Ahrens „Verborgene Ordnung" und das Menzerathsche Gesetz. In M. Faust, R. Harweg u.a. (Hg.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik. Festschrift für Peter Hartmann,* Tübinger Beiträge zur Linguistik, Bd. 215, (S 31-39), Tübingen: Narr.

Altmann, G. (1985). Sprachtheorie und mathematische Modelle. *SAIS Arbeitsberichte,* Heft 8, 1-13.

Altmann, G. (1987). The Levels of Linguistic Investigation. *Theoretical Linguistics,* 14, 227-239.

Altmann, G. (1988). Verteilungen der Satzlängen. In K.P. Schulz (Hg.), *Glottometrika 9* (S. 147-169), Bochum: Brockmeyer.

Altmann, G. (1992a). Das Problem der Datenhomogenität. In B. Rieger (Hg.), *Glottometrika 13* (S. 287-298), Bochum: Brockmeyer.

Altmann, G. (1992b). Sherman's Laws of Sentence Length Distribution. In P. Saukkonen (Hg.), *What is language synergetics? Acta Universitatis Ouluensis, Series B, Humaniora 16* (S. 38-39), Oulu.

Altmann, G. (1993). Science and Linguistics. In R. Köhler & B. Rieger (Hg.), *Contributions to Quantitative Linguistics* (S.3-10), Dordrecht: Kluwer.

Altmann, G.: Privater Briefwechsel vom 11. Februar 1994.

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In P. Schmidt (Hg.), *Glottometrika 15* (S. 166-180), Trier: WVT.

Altmann, G., & Schwibbe, M. H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Altmann, H. (1981). *Formen der „Herausstellung" im Deutschen. Rechtsversetzung, Linksversetzung, Freies Thema und verwandte Konstruktionen.* Tübingen: Niemeyer (Linguistische Arbeiten).

Bamberger, R., & Vanecek, E. (1984). *Lesen - Verstehen - Lernen - Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache*. Wien: Jugend und Volk, Frankfurt: Diesterweg.

Beneš, E. (1976). Syntaktische Besonderheiten der deutschen wissenschaftlichen Fachsprache. In K.-H. Bausch, W.H.U. Schewe & H.-R. Spiegel (Hg.), *Fachsprachen* (S. 88-98), Berlin, Köln: Beuth.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Betten, A. (1976). Ellipsen, Anakoluthe und Parenthesen. *Deutsche Sprache*, 4, 207-230.

Brass, W. (1958). Simplified Methods of Fitting the Truncated Negative Binomial Distribution. *Biometrika,* 45, 59-68.

Brinkmann, H. (1974). Reduktion in gesprochener und geschriebener Rede. *Gesprochene Sprache. Jahrbuch 1972 des Instituts für deutsche Sprache* (S. 144-162), Düsseldorf.

Bronstein, I.N., & Semendjajew, K.A. (1989). *Taschenbuch der Mathematik.* 24. Aufl. hrsg. von Grosche, G., Ziegler, V. & Ziegler, D., Frankfurt/M.: Deutsch.

Buch, K.R. (1969). A note on sentence-length as random variable. In L. Dolezel & R.W. Bailey (Hg.), *Statistics and Style* (S. 76-79), New York: Elsevier.

Bunge, M. (1967). *Scientific Research. Vol 1.* Berlin: Springer.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax.* 2. Überarb. Aufl., Frankfurt/M: Athenäum.

Bußmann, H. (1990). *Lexikon der Sprachwissenschaft.* 2. Völlig neu bearb. Aufl.. Stuttgart: Kröner.

Cherubim, D. (1993). Variatio delectat. Oder: Dabeisein ist alles. In W.P. Klein & I. Paul (Hg.), *Sprachliche Aufmerksamkeit* (S. 29-34), Heidelberg: Winter.

Clayman, D.L. (1981). Sentence length in Greek hexameter poetry. In R. Grotjahn (Hg.), *Hexameter Studies* (S. 107-136), Bochum: Brockmeyer.

Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit-tests. *Journal of the Royal Statistical Society,* B 46, 440-464.

David, F.N., & Johnson, N.L. (1952). The truncated Poisson. *Biometrics,* 8, 275-282.

Farr, J. N., Jenkins, J.J., & Peterson, D. G. (1951). Simplification of Flesch Reading Ease Formula. *Journal of Applied Psychology,* 35, 333-337.

Flesch, R.A. (1948). A New Readability Yardstick. *Journal of Applied Psychology,* 32, 221-233.

Fucks, W. (1968). *Nach allen Regeln der Kunst.* Stuttgart: Dt. Verl.-Anst.

Fucks, W., & Lauter, J. (1965). Mathematische Analyse des literarischen Stils. In H. Kreuzer & R. Gunzenhäuser (Hg.), *Mathematik und Dichtung* (S. 107-122), München: Nymphenburger Verlag.

Gerlach, R. (1982). Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie. In W. Lehfeldt & U. Strauß (Hg.), *Glottometrika 4* (S. 95-102). Bochum: Brockmeyer.

Groeben, N. (1982). *Leserpsychologie: Textverständnis - Textverständlichkeit.* Münster: Aschendorff.

Grotjahn, R. (1982). Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft,* 1, 44-75.

Grotjahn, R., & Altmann, G. (1993). Modelling the distribution of word length: Some methodological problems. In R. Köhler & B. Rieger (Hg.), *Contributions to Quantitative Linguistics,* (S. 141-153), Dordrecht: Kluwer.

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. In L. Hřebíček (Hg.), *Glottometrika 11* (S. 142-156), Bochum: Brockmeyer.

Heups, G. (1980). Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. (schriftliche

Hausarbeit im Rahmen der fachwissenschaftlichen Prüfung für das Lehramt an Gymnasien). Göttingen.

Johnson, N.L., & Kotz, S. (1969). *Discrete distribution.* New York: Houghton Mifflin.

Kemp, K.W. (1976). Personal observations on the use of statistical methods in quantitative linguistics. In A. Jones & R.F. Churchhouse (Hg.), *The computer in literary and linguistic studies* (S. 59-77), Cardiff: University of Wales Press.

Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. In W. Lehfeldt & U. Strauß (Hg.), *Glottometrika 4* (S. 103-113), Bochum: Brockmeyer.

Köhler, R. (1987). System Theoretical Linguistics. *Theoretical Linguistics,* 14, Nr. 2/3, 241-257.

Köhler, R. (1990). Elemente der synergetischen Linguistik. In R. Hammerl, *Glottometrika 12* (S. 179-187), Bochum: Brockmeyer.

Köhler, R., & Altmann, G. (1986). Synergetische Aspekte der Linguistik. *Zeitschrift für Sprachwissenschaft,* 5, 253-265.

Köhler, R., & Altmann, G. (Manuskript). *Einladung zur quantitativen Linguistik* (Arbeitstitel)(1-18), Kapitel 2 des geplanten Buches.

Krengel, U. (1988). *Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Braunschweig: Vieweg.

Kürschner, W. (1993). *Grammatisches Kompendium. Systematisches Verzeichnis grammatischer Grundbegriffe.* 2. überarb. u. stark erw. Aufl.. Tübingen: Francke.

Lewandowski, T. (1990). *Linguistisches Wörterbuch.* 5. überarb. Aufl., Bd. 1-3. Heidelberg, Wiesbaden: Quelle und Meyer.

Lüger, H.-H. (1983). *Pressesprache.* Tübingen: Niemeyer (Germanistische Arbeitshefte 28).

Lutz, B. (Hg.) (1989). *Metzler-Philosophen-Lexikon: 300 biographisch-werkgeschichtliche Porträts von den Vorsokratikern bis zu den neuen Philosophen.* Unter redaktioneller Mitarbeit von C. Dehlinger u.a. Stuttgart: Metzler.

Maier, K.-E. (1980). *Jugendliteratur. Formen, Inhalte, pädagogische Bedeutung.* 8., neu bearb. Aufl. von „Jugendschrifttum". Bad Heilbrunn/Obb: Klinkhardt.

Michelsen, P. (1990). Literatur und Sprache. In G. Buhr, F.A. Kittler & H. Turk (Hg.), *Das Subjekt der Dichtung. Festschrift für Gerhard Kaiser* (S. 141-152), Würzburg: Königshausen & Neumann.

Mihm, A. (1973). Sprachstatistische Kriterien zur Tauglichkeit von Lesebüchern. *Linguistik und Didaktik,* 4, 117-127.

Mistrík, J. (1971/72). Exakte Methoden einer Texttypologie. *Zeitschrift für Slavische Philologie,* 36, 318-331.

Mistrik, J. (1973). *Exakte Typologie von Texten*. München: Sagner in Komm (Arbeiten und Texte zur Slavistik).

Moore, D., & Spruill, M.C. (1975). Unified large-sample theory of general chisquared statistics for tests of fit. *Annals of Statistics*, 3, 599-616.

Morton, A.Q. (1965). The authorship of Greek prose. *Journal of the Royal Statistical Society*, A 128, 169-233.

Morton, A.Q., & Mc Leman, J. (1966). *Paul, the man and the myth*. London: Hoder and Stoughton.

Mosteller, F., & Wallace, D. L. (1963). Interference in an authorship problem. *Journal of the American Statistical Association*, 58, 275-309.

Pieper, U. (1979). *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse*. Tübingen: Narr (Ars linguistica 5).

Riesel, E. (1963). *Stilistik der deutschen Sprache*. 2., durchges. Aufl.. Moskau: Staatsverlag „Hochschule".

Sampford, M.R. (1955). The truncated negative binomial distribution. *Biometrika*, 42, 58-69,

Schindler, W. (1990). *Untersuchung zur Grammatik appositionsverdächtiger Einheiten im Deutschen*. Tübingen: Niemeyer.

Schmidt, F. (1966). *Zeichen und Wirklichkeit*. Stuttgart: Kohlhammer.

Sherman, L.A. (1888). Some observations upon the sentence-length in English prose. *University of Nebraska Studies*, 1, 119-130.

Sowinski, B. (1988). *Deutsche Stilistik*. Überarb. im Sept. 1978. Frankfurt/M: Fischer Taschenbuch Verlag.

Sichel, H.S. (1974). On a distribution representing sentence-length in prose. *Journal of the Royal Statistical Society*, A 120, 331-346.

Spang-Hanssen, H. (1963). *Sentence-length and statistical linguistics*. In *Structures and Quanta* (S. 58-72), Copenhagen.

SPIEGEL: Telefonat mit der SPIEGEL-Redaktion vom 26. April 1994.

Stegmüller, W. (1993). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Bd. 1: Erklärung, Begründung, Kausalität*. 2., verb. und erw. Aufl.. Berlin, Heidelberg, New York: Springer.

Tallentire, D.R. (1976). Confirming intuitions about style using concordances. In A. Jones & R.F. Churchhouse (Hg.), *The computer in literary and linguistic studies* (S. 309-328), Cardiff: University of Wales Press.

Wake, W.C. (1957). Sentence-length distribution of Greek authors. *Journal of the Royal Statistical Society*, A 120, 331-346.

Willams, C.B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, 31, 356-361.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Yule, G.U. (1939). On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika*, 30, 363-390.

Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge: University Press.

Glottometrika 16, 1997, 276-285

Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart

Karl-Heinz Best

1. Die unterschiedliche Häufigkeit von Wortarten in Texten verschiedener Sprachen und Funktionalstile ist in der letzten Zeit wieder in den Vordergrund des Interesses der Forschung getreten (Hammerl, 1990; Schweers & Zhu, 1991). Bei diesen Arbeiten geht es um die Frage, welchen Gesetzen die Wortartenhäufigkeiten in Texten womöglich folgen könnten. Dabei sind unterschiedliche Hypothesen geprüft worden, meist mit guten Ergebnissen. So hat Hammerl es unternommen, insgesamt 110 russische und deutsche Texte bzw. Textfragmente daraufhin zu testen, ob sie der sog. Zipf-Dolinski-Verteilung entsprechen. Es zeigte sich, daß 107 Text(fragment)e dem Modell folgen und nur drei ihm nicht gehorchen. Die Vermutung, das Versagen des Modells in den drei genannten Fällen könne darauf zurückzuführen sein, daß diese Texte länger als die übrigen sind, nämlich über 2000 Wörter lang, bestätigte sich bei der Überprüfung von drei weiteren „überlangen" Texten: Auch diese folgten dem Modell nicht. Man kann aus diesen Ergebnissen mit Hammerl den Schluß ziehen, daß längere Texte einen erhöhten Grad an Inhomogenität der Texteigenschaften aufweisen und deshalb die vermuteten Gesetze, denen die Wortartenhäufigkeiten folgen, nur an entsprechend kurzen Texten relativ ungestört zutage treten. Der Grund für das Versagen des Modells kann allerdings auch darin bestehen, daß Hammerl zu seiner Überprüfung den Chiquadrattest verwendet (Hammerl, 1990:149ff.), der die unangenehme Eigenschaft hat, daß er bei wachsendem Stichprobenumfang bei gleichbleibenden Freiheitsgraden wachsende X^2 - Werte liefert, die dann zur Ablehnung der Modelle führen. (vgl. Grotjahn & Altmann, 1993) Aus diesem Grund wird in der vorliegenden Untersuchung statt des Chiquadrattests der Determinationskoeffizient als Prüfkriterium verwendet.

Die Untersuchung von Schweers & Zhu bestätigte die Ergebnisse Hammerls für einen lateinischen und zwei deutsche Texte; für den chinesischen Text konnte nur ein anderes Modell, die negative hypergeometrische Verteilung, mit Erfolg angepaßt werden. Dieses Modell hat aber den Vorteil, daß es auch den lateinischen und die beiden deutschen Texte korrekt darstellen kann. Bei diesem Stand der Untersuchungen liegt nun die Vermutung nahe, daß es sinnvoll sein könnte, noch nach weiteren Hypothesen für die Wortartenhäufigkeiten in Texten zu suchen.

2. Die vorliegende Untersuchung dient genau diesem Zweck. Um Problemen mit der möglichen Inhomogenität längerer Text aus dem Wege zu gehen, wurden nur kurze Texte aus einem einzigen Funktionalstil, dem der Prosaliteratur der Gegenwart, ausgewertet. Die Texte sind um 1000 Wörter lang und bleiben damit deutlich unter der kritischen Grenze von ca. 2000 Wörtern Textlänge. So kann man erwarten, daß die Inhomogenität der Daten zumindest innerhalb der einzelnen Texte nicht allzu groß sein wird.

3. Statt der von Hammerl und Schweers & Zhu herangezogenen Modelle wird hier ein weiteres überprüft, das Altmann (1993) entwickelt hat. Es handelt sich dabei um ein Modell für beliebige sprachliche Rangordnungen, das nicht als Wahrscheinlichkeitsfunktion, sondern als gewöhnliche Funktion dargestellt und an die relativen Häufigkeiten angepaßt wird:

$$(1) \qquad y_x = \frac{\binom{b+x}{x-1}}{\binom{a+x}{x-1}} y_1, \qquad x=1,2,\ldots,k$$

Die Güte der Anpassung wird mit dem Determinationskoeffizienten geprüft:

$$(2) \qquad D = 1 - \frac{\sum_{x=1}^{k}(y_x - \hat{y}_x)^2}{\sum_{x=1}^{k}(y_x - \bar{y})^2}$$

Je höher der Wert für D, desto besser ist die Übereinstimmung der Daten mit dem Modell. Dieses Verfahren hat den Vorteil, daß es die o.a. Probleme, die bei der Überprüfung der Anpassung mit Hilfe des Chiquadrattests entstehen, vermeidet und keine Rücksicht auf Freiheitsgrade zu nehmen braucht.

4. Als „Wort" werden „graphematische" Wörter (Bünting & Bergenholtz, 1989:36f, 39) gewertet, also zusammenhängende Buchstabenfolgen, die von anderen durch Leerstellen im Druckbild oder durch Interpunktionszeichen getrennt erscheinen. Zusätzlich gilt: Bindestrichkomposita werden als ein Wort aufgefaßt, ebenso Verbformen mit integrierter Infinitivkonjunktion „zu" und Kontaminationen mit Präpositionen (im, am etc.). Wörter mit Abkürzungspunkt (e.V.) und Zahlen (45) werden ausbuchstabiert (hier: „eingetragener Verein"; „fünfundvierzig") und entsprechend dieser Auflösung die Zahl der Wörter und die Wortarten bestimmt.

Die Wortartbestimmung erfolgt im Wesentlichen nach der Variante A des Modells von Flämig (1981:491). Abgetrennte Präfixe (stellt ... auf) werden als Adverbien, abgetrennte Kompositionsbestandteile (stellt ... fest) als Adjektiv etc. aufgefaßt. Kontaminationen wie „im" gelten als Präpositionen.

Ein gewisses Problem stellen die Interjektionen dar. In den untersuchten Texten kamen entweder keine oder nur jeweils eine pro Text vor. Der Verlauf der Kurve bzw. das Resultat der Anpassung kann aber durch den steilen Sprung vom vorletzten zum letzten Rang der Wortarten stark beeinträchtigt werden. In solchen Fällen empfiehlt es sich, a) die Interjektion ganz außer Acht zu lassen oder b) ihre Häufigkeit dem vorletzten Rang zuzuschlagen, was ein durchaus übliches Verfahren darstellt. In dieser Untersuchung wurde bei Text 1 zum Zwecke der Illustration die Interjektion beibehalten, in den Texten 4 und 8 jedoch eliminiert. Letzteres entspricht auch Flämigs Modellvariante A (Flämig, 1981:491). eine zusätzliche Rechtfertigung für diese Entscheidung läßt sich noch daraus ableiten, daß Interjektionen einen besonderen Status im Wortartensystem beanspruchen, der „auf allen Ebenen der grammatischen Beschreibung sichtbar" (Fries, 1992:311) wird.

Im Gegensatz zu Flämig, der zwar die Wortart „Artikel" in seinem Modell nicht berücksichtigt, wohl aber „artikelfähig" als Kriterium für Substantive benötigt, wird hier „Artikel" als eigene Wortart aufgefaßt. Nur bestimmter und unbestimmter Artikel, jedoch keine sonstigen „Artikelwörter" werden dieser Wortart zugerechnet. (Zu einigen weiteren Aspekten des Modells von Flämig vgl. Best, 1980:27ff; Ossner, 1989:98ff)

5. Die Ergebnisse der Untersuchung sind in den Tabellen 1 bis 10 dargestellt. Hierbei bedeutet n_x die absolute Häufigkeit der Wortart, y_x die relative Häufigkeit und $\hat{y}_x$ die berechneten Werte.

Man erkennt an den durchwegs guten Werten für den Koeffizienten D, daß in allen Fällen eine hohe Übereinstimmung zwischen dem Modell und den Beobachtungen an den einzelnen Texten erzielt werden konnte. Das von Altmann vorgeschlagene Modell für beliebige Rangordnungen hat sich damit im Fall der Wortartenhäufigkeiten bewährt.

Man sieht aber auch, daß die Parameter a und b beträchtliche Unterschiede aufweisen. Der Grund dafür besteht darin, daß alle Rechnungen mit einem Optimierungsverfahren durchgeführt wurden, das im Unterschied zu klassischen Punktschätzern die Parameter im Laufe des Prozesses solange ändert, bis die beste Anpassung erreicht wird. Man darf diese Parameter daher nicht als Textcharakteristika betrachten.

Tabelle 1

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Verb	1	313	24.32	24.32
Pronomen	2	262	20.36	18.62
Adverb	3	193	15.00	14.39
Substantiv	4	163	12.67	11.21
Konjunktion	5	150	11.66	8.81
Artikel	6	104	8.08	6.97
Adjektiv	7	56	4.35	5.56
Präposition	8	45	3.50	4.46
Interjektion	9	1	0.08	3.60
$a = 30.1447$; $b = 22.6085$; $D = 0.9421$				

Text 1: P. Bichsel, Der Mann, der nichts mehr wissen wollte (Kindergeschichte)

Tabelle 2

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Substantiv	1	229	21.91	21.91
Verb	2	172	16.46	16.55
Artikel	3	144	13.78	13.60
Pronomen	4	132	12.63	11.70
Adjektiv	5	120	11.48	10.35
Adverb	6	89	8.52	9.33
Präposition	7	80	7.66	8.54
Konjunktion	8	79	7.56	7.89
$a = 0.6713$; $b = 0.0174$; $D = 0.9781$				

Text 2: P. Bichsel, Und sie dürfen sagen, was sie wollen

Tabelle 3

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Substantiv	1	238	19.52	19.52
Verb	2	168	13.78	15.73
Adverb	3	167	13.70	13.55
Artikel	4	158	12.96	12.09
Pronomen	5	155	12.72	11.02
Präposition	6	125	10.25	10.20
Adjektiv	7	110	9.02	9.54
Konjunktion	8	98	8.04	9.00
$a = 0.4952$; $b = 0.0107$; $D = 0.9035$				

Text 3: J. Bobrowski, Betrachtung eines Bildes

Tabelle 4

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Verb	1	192	20.56	20.56
Adverb	2	174	18.63	16.95
Substantiv	3	150	16.06	14.14
Pronomen	4	141	15.10	11.92
Artikel	5	88	9.42	10.15
Konjunktion	6	79	8.46	8.71
Adjektiv	7	67	7.17	7.53
Präposition	8	42	4.50	6.56
$a = 14.5375$; $b = 11.6300$; $D = 0.9100$				

Text 4: J. Bobrowski, Mäusefest

Tabelle 5

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Substantiv	1	274	22.51	22.51
Verb	2	164	13.48	16.46
Adverb	3	161	13.23	13.51
Artikel	4	142	11.67	11.69
Adjektiv	5	140	11.50	10.43
Pronomen	6	138	11.34	9.49
Präposition	7	109	8.96	8.76
Konjunktion	8	89	7.31	8.17
$a = 0.0075$; $b = -0.5319$; $D = 0.9008$				

Text 5: J. Bobrowski, Rainfarn

Tabelle 6

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Substantiv	1	221	23.54	23.54
Verb	2	162	17.25	17.14
Adverb	3	125	13.31	13.69
Artikel	4	115	12.25	11.51
Pronomen	5	92	9.80	9.98
Präposition	6	83	8.84	8.85
Konjunktion	7	75	7.99	7.98
Adjektiv	8	66	7.03	7.28
$a = 0.8522$; $b = 0.0772$; $D = 0.9962$				

Text 6: G. de Bruyn, Stallschreiberstraße 45

Tabelle 7

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Verb	1	249	21.54	21.54
Pronomen	2	193	16.70	17.72
Substantiv	3	176	15.22	14.74
Adverb	4	154	13.32	12.39
Adjektiv	5	108	9.34	10.50
Konjunktion	6	108	9.34	8.98
Artikel	7	105	9.08	7.74
Präposition	8	63	5.45	6.71
$a = 15.8138$; $b = 12.6510$; $D = 0.9629$				

Text 7: G. de Bruyn, Vergißmeinnicht

Tabelle 8

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Substantiv	1	261	23.64	23.64
Verb	2	185	16.76	17.14
Artikel	3	148	13.41	13.68
Pronomen	4	139	12.59	11.49
Adjektiv	5	121	10.96	9.97
Präposition	6	99	8.97	8.85
Adverb	7	85	7.70	7.98
Konjunktion	8	65	5.89	7.28
$a = 0.7818$; $b = 0.0173$; $D = 0.9802$				

Text 8: G. Kunert, Die Taucher

Tabelle 9

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Substantiv	1	169	18.15	18.15
Verb	2	166	17.83	15.96
Pronomen	3	159	17.08	14.08
Adjektiv	4	112	12.03	12.45
Konjunktion	5	92	9.88	11.04
Artikel	6	90	9.67	9.82
Adverb	7	85	9.13	8.75
Präposition	8	68	7.30	7.82
$a = 43.8065$; $b = 38.2850$; $D = 0.8938$				

Text 9: G. Kunert, Warum schreiben

Tabelle 10

Wortart	Rang	n_x	y_x	$\hat{y}_x$
Verb	1	168	20.24	20.24
Substantiv	2	151	18.19	16.85
Pronomen	3	121	14.58	14.24
Adverb	4	105	12.65	12.19
Artikel	5	85	10.24	10.54
Konjunktion	6	79	9.52	9.21
Adjektiv	7	74	8.92	8.11
Präposition	8	47	5.66	7.20
$a = 10.3237$; $b = 8.2590$; $D = 0.9686$				

Text 10: P. Bichsel, Lesebuchgeschichte (ohne schweizerdeutsche Einschübe)

6. Um die gewonnenen Daten auch für stilistische Vergleiche leicht nutzen zu können, werden sie noch in zwei Tabellen zusammengestellt (vgl. Tabelle 11 und 12), die eine Übersicht über die Häufigkeitsränge und die relativen Anteile der Wortarten in den untersuchten Texten darbieten.

Tab. 1: Häufigkeitsränge der Wortarten in den zehn Texten

Text	Subst.	Verb	Adj.	Adv.	Art.	Pron.	Präp.	Konj
1	4	1	7	3	6	2	8	5
2	1	2	5	6	3	4	7	8
3	1	2	7	3	4	5	6	8
4	3	1	7	2	5	4	8	6
5	1	2	5	3	4	6	7	8
6	1	2	8	3	4	5	6	7
7	3	1	5	4	7	2	8	6
8	1	2	5	7	3	4	6	8
9	1	2	4	7	6	3	8	5
10	2	1	7	4	5	3	8	6

Tab. 2: Relativer Anteil der Wortarten (ohne Interjektionen) in den zehn Texten (in Prozent)

Text	Subst	Verb	Adj.	Adv.	Art.	Pron.	Präp.	Konj.
1	12.67	24.32	4.35	15.00	8.08	20.36	3.50	11.66
2	21.91	16.46	11.48	8.52	13.78	12.63	7.66	7.56
3	19.52	13.78	9.02	13.70	12.96	12.72	10.25	8.04
4	16.06	20.56	7.17	18.63	9.42	15.10	4.50	8.46
5	22.51	13.48	11.50	13.23	11.67	11.34	8.96	7.31
6	23.54	17.25	7.03	13.31	12.25	9.80	8.84	7.99
7	15.22	21.54	9.34	13.32	9.08	16.70	5.45	9.34
8	23.64	16.76	10.96	7.70	13.41	12.59	8.97	5.89
9	18.15	17.83	12.03	9.13	9.67	17.08	7.30	9.88
10	18.19	20.24	8.92	12.65	10.24	14.58	5.66	9.52

Beide Tabellen zeigen selbst in diesem kleinen, relativ homogenen Korpus beträchtliche Unterschiede zwischen den untersuchten Texten. Es läßt sich aber zeigen, daß die Ränge recht gut übereinstimmen, d.h. die Verschiebung der Ränge in dieser Gruppe von Texten ist lediglich eine nichtsignifikante Fluktuation. Der Kendallsche Konkordanzkoeffizient für mehrere Rangierungen

(vgl. z.B. Gibbons, 1971) ergibt W = 0.73 und X^2 = 51.17 mit 7 Freiheitsgraden, was eine hohe Übereinstimmung der Ränge bedeutet. Die untersuchten Texte finden sich in folgenden Ausgaben:

Peter Bichsel. *Stockwerke*. Stuttgart: Reclam 1986
Johannes Bobrowski. *Lipmanns Leib*. Stuttgart: Reclam 1987
Günter de Bruyn. *Babylon*. Frankfurt: Fischer Taschenbuch Verlag 1992
Günter Kunert. *Der Hai*. Stuttgart: Reclam 1990

Literatur:

Altmann, G. (1993). Phoneme counts. In G. Altmann (Hg.), *Glottometrika 14* (S. 54-68), Trier: WVT.

Best, K.-H. (1980). Überlegungen zu einigen Problemen morphologisch orientierter Wortartenmodelle. *Kwartalnik Neofilologiczny*, 27, 23-42.

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II* (S. 19-30), Stuttgart: Steiner.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax*. Frankfurt: Athenäum.

Flämig, W. (1981). Das Wort. In K.E. Heidolph (Hg.) u.a., *Grundzüge einer deutschen Grammatik* (S. 458-496), Berlin: Akademie.

Fries, N. (1992). Interjektionen, Interjektionsphrasen und Saztmodus. In I. Rosengren (Hg.), *Satz und Illokution. Bd. 1.* (S. 307-341), Thübingen: Niemeyer.

Grotjahn, R., & Altmann, G. (1993). Modelling the Distribution of Word Length. In R. Köhler & B.B. Rieger (Hg.), *Contributions to quantitative linguistics (QUALICO, Trier 1991)* (S. 141-153), Dordrecht: Kluwer.

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. In L. Hřebíček (Hg), *Glottometrika 11* (S. 142-156), Bochum: Brockmeyer.

Ossner, J. (1989). Wortarten: Form- und Funktionsklassen. *Zeitschrift für Literaturwissenschaft und Linguistik*, 19, 94-117.

Schweers, A., & Zhu, J. (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In U. Rothe (Hg.), Diversification Processes in Language: Grammar (S. 157-165), Hagen: Margit Rottmann Medienverlag.

Autorenadressen

Stefan Ammermann, Karolinenweg 20/202, D 37075 Göttingen
Claudia Balschun, Theodor-Heuss-Str. 11/003, D-37075 Göttingen
Olaf Bartels, Wallstr. 5, D-21682 Stade
Hans-Hermann Bartens, Wartburgweg 2b, D-37085 Göttingen
Gabi Behrmann, Zimmermannstr. 62/205, D-37075 Göttingen
Karl-Heinz Best, Im Siebigsfeldt 17, D-37115 Duderstadt
Birte Christiansen, Annastr. 64/73, D-37083 Göttingen
Alice Hasse, Otto-Wels-Weg 1a, D-37077 Göttingen
Martina Hein, c/o Münzer, Kreuzbergring 56 A 01/02, D-37075 Göttingen
Cecilie Hollberg, Lenaustr. 14-15, D-12047 Berlin
Marianne Janssen, Groner Landstr. 9, App. 268, D-37073 Göttingen
Saskia Kuhr, Eisenacherstr. 7, D-37085 Göttingen
Brigitta Niehaus, Beuthener Weg 2, D-30916 Isernhagen
Esther Rauhaus, Tilsiter Str. 4, D-37083 Göttingen
Gesa Riedemann, Hagenbeckstr. 60, Haus 3, Zi. 212. D-22527 Hamburg
Winfred Röttger, Robert-Koch-Str. 2, D-37075 Göttingen
Ludmila Uhlířová, CSc., Ústav pro Jasyk český AV ČR, Letenská 4, Praha 1, Tschechische Republik.