Glottometrics 8, 2004, 23-41

Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable

(am Beispiel slowenischer Texte)

Emmerich Kelih, Peter Grzybek
Universität Graz

Abstract: The present study is a contribution to the study of sentence length. Specifically, the study focuses on the question of factors influencing the theoretical modeling of frequency distributions of sentence lengths. Slovenian texts are analyzed on three analytical levels: individual texts, complex texts, and a text corpus. On the basis of this material, the impact of a broadly accepted smoothing procedure (smoothing by forming specific intervals) on the adequacy of theoretical models is controlled.

Keywords: Sentence length, Slovenian

1. Theoretische Modellierung der Satzlängenverteilung:

Die theoretische Modellierung der Satzlänge – d.h. die Frage, ob die Satzlängenverteilung durch ein entsprechendes theoretisches Verteilungsmodell beschrieben werden kann – hat eine mehr als 50-jährige Geschichte. Fast ein halbes Jahrhundert nach den ersten Grundsatzüberlegungen von Sherman (1888) warf Yule (1939) in seiner Untersuchung zur Satzlänge in englischen Texten erstmals diese Frage auf; dabei nannte er zwar explizit kein theoretisches Verteilungsmodell, wies aber auf folgende Beobachtung hin: „They are not of the Poisson type but of the type in which the square of the standard deviation largely exceeds the mean" (Yule 1939: 371). In Anlehnung an diese Untersuchung schlug wenig später dann Williams (1940) vor, dass man bei der Verteilung der Satzlänge nicht die absolute Anzahl der Wörter pro Satz als Variable bestimmen solle, sondern die jeweilige logarithmisch transformierte Wortanzahl pro Satz. Die von Williams (1940) durchgeführte Re-Analyse der Daten von Yule beschränkte sich auf eine graphische Darstellung; diese zeigte seiner Ansicht nach, dass die Häufigkeitsverteilung der *x*-silbigen Wörter einer Normalverteilung folgt. Aufgrund der empirischen Untersuchung von drei Texten postulierte er sodann die genannte Lognormal-Verteilung als allgemein gültiges theoretisches Modell der Satzlängenverteilung (vgl. Williams 1940: 360).

Eine Kritik erfuhr der Ansatz von Williams erst Jahrzehnte später durch Sichel (1974), der auf einige Unzulänglichkeiten des Vorgangsweise hinweist. So kritisierte Sichel zu Recht, dass die Lognormal-Verteilung nur aufgrund einer graphischen Darstellung der Verteilung der Satzlängen postuliert wurde, ohne dass die erwähnte Verteilung einem statistischen Prüfverfahren unterzogen worden wäre, welches den Grad der Adäquatheit zwischen theoretischem Modell und empirischer Beobachtung hätte erbringen können. Sichel (1974) selbst schlug seinerseits eine zusammengesetzte Poisson-Verteilung als allgemeines Modell der Satzlängenverteilung vor, bei welcher ein Parameter wiederum als Zufallsvariable einer weiteren Verteilung folgt (vgl. Sichel 1974: 26f.). Diese Distribution wurde von ihm an acht lateini-

schen, griechischen und englischen Texten inklusive entsprechender statistischer Tests überprüft. Diese ersten Überlegungen zu einer theoretischen Modellierung wurden dann in weiterer Folge in einem allgemeinen Ansatz von Altmann aufgegriffen, der damit – wie darzulegen sein wird – eine neue Perspektive in dieser Diskussion aufzeigte.

1.1. Neuansatz in der theoretischen Modellierung der Satzlängenverteilung

Die oben einleitende dargestellten Arbeiten zur theoretischen Modellierung von Satzlängen wurden von Altmann (1988a) einer allgemeinen Kritik unterzogen: dabei wurde seinerseits darauf verwiesen, dass das Auffinden einer Verteilung (hier bezogen auf die zusammengesetzte Poisson-Verteilung von Sichel) in der Regel inhaltlich (d.h. hier: linguistisch) schwer zu begründen bzw. zu interpretieren ist und somit für die Analyse von sprachlichen Phänomenen keine nennenswerten Erkenntnisse beisteuert. Dem stellte Altmann (1988b) einen grundlegenden Neuansatz gegenüber, der darin besteht, die Distribution von sprachlichen Einheiten in einen synergetisch-linguistischen Kontext (vgl. Köhler 1986) zu stellen. Bezogen auf die Frage der Verteilung von Satzlängen werden a priori folgende systeminterne und systemexterne Faktoren als Einflussfaktoren in Betracht gezogen (Altmann (1988a: 152):

a – Wirkung des Produzenten (Stil u.a)
b – Wirkung des Rezipienten (der Sprachgemeinschaft, Rücksicht auf den Hörer)
c – Faktoren des Textes
d – Faktoren der Ebene

Ausgehend von zwei Annahmen, nämlich
1. dass jegliche Verteilung der Längen in einem Text (unter anderem, aber nicht ausschließlich also auch der Satzlängen) gesetzmäßig organisiert ist, und
2. dass es ausreichend ist, Annahmen über die Differenz zweier jeweils benachbarter Wahrscheinlichkeiten zu machen,

stellte Altmann (1988b) den folgenden Ansatz auf: Sei die Differenz benachbarter Klassen

$$(1) \qquad P_x - P_{x-1} = \Delta P_{x-1}.$$

und diese Differenz sei nicht konstant, sondern hänge vom jeweiligen P_{x-1} ab, so dass sich der Quotient

$$D = \frac{P_x - P_{x-1}}{P_{x-1}} = \frac{\Delta P_{x-1}}{P_{x-1}}$$

ergibt. Im Hinblick auf die Berechnung der Satzlänge in Anzahl der Teilsätze kommt der Faktor *d* folglich nicht zum Tragen; somit ergibt sich nach Altmann (1988b) die folgende Gleichung:

$$(2) \qquad D = \frac{b - ax}{cx}$$

Hierbei wirken die der Zipf'schen Kräfte von Unifikation bzw. Diversifikation *a* und *b* „gestaltend", während *c* „bremsend" wirkt (*a* hat ein negatives Vorzeichen, insofern der

Produzent versucht, seinen eigenen Stil in die Bindung der Textsorte einzubringen); weiterhin wirkt *b* „global“, während *a* und *c* „lokal“ wirken. Falls allerdings die Satzlänge in Wortanzahl gemessen wird, kommt der intervenierende Faktor der Sprachebene (*d*) hinzu, was zu der Quotienten (3) führt:

$$(3) \qquad D = \frac{b - ax}{cx + d}.$$

Dies führt zu den beiden Ansätzen (4) und (5)

$$(4) \qquad P_x = \frac{b - ax}{cx} P_{x-1}$$

$$(5) \qquad P_x = \frac{b - ax}{cx + d} P_{x-1}$$

Wir können uns hier die weiteren detaillierten Ableitungen ersparen und wollen uns statt dessen auf die aus ihnen (nach Reparametrisierung) resultierenden Verteilungsmodelle beschränken. So wird für die Satzlänge, gemessen in der Anzahl der Teilsätze (clauses) pro Satz, die *negative Binomialverteilung* (6) als adäquates Verteilungsmodell hergeleitet. Falls die Satzlänge in Anzahl der Worte, nicht der Anzahl der clauses pro Satz gemessen wird, kommt die Wirkung der intervenierenden Ebene des Teilsatzes (*d*) hinzu; unter dieser Bedingung wird die *Hyperpascal-Verteilung* (7) als adäquates Modell der Satzlängenverteilung postuliert.

$$(6) \qquad P_x = \binom{r + x - 1}{x} p^r q^x \qquad x = 0, 1, 2, \ldots$$

$$(7) \qquad P_x = \frac{\binom{k + x - 1}{x}}{\binom{m + x - 1}{x}} q^x P_0 \qquad x = 0, 1, 2, \ldots$$

Die Gesetzeshypothese zur Satz/Wort-Variante wurde von Altmann (1988a) an 245 Texten des Altgriechischen, Englischen und Slowakischen überprüft, wobei nur in neun Fällen keine signifikante Übereinstimmung mit dem Modell erzielt wurde. Bei den wenigen Texten, die keine Übereinstimmung mit der Hyperpascal-Verteilung zeigten, handelte es sich um solche Texte mit ungeklärter Autorschaft beziehungsweise Texte, an denen seitens der Editoren Modifikationen durchgeführt worden waren. In den anderen Fällen handelte es sich um zusammengesetzte Stichproben, die „nur unter sehr günstigen Umständen dem Gesetz folgen” (Altmann 1988a: 159). Durch die Überprüfung der clause-Variante in zehn Texten unterschiedlicher Sprachen konnte die Hypothese, dass die Satzlängenhäufigkeitsverteilung durch die negative Binomialverteilung modelliert werden kann, bestätigt werden.

Während sowohl die einleitende dargestellten Untersuchungen als auch die zuletzt referierten Überlegungen von Altmann zur theoretischen Modellierung der Häufigkeit von

Satzlängen durch theoretische Wahrscheinlichkeitsverteilungen von der sprachübergreifenden Relevanz der jeweiligen Konzepte ausgingen, haben jüngere Untersuchungen in diesem Bereich (Niehaus 2001, Best 2001 Grzybek 1999 u.a.) es als wahrscheinlicher erscheinen lassen, dass eher von verschiedenen Modellen auszugehen ist, die sich zwar vermutlich auf einen allgemeinen Ansatz wie den von Wimmer/Altmann (1996) zurückführen lassen, von denen wir aber bislang nicht wissen, unter welchen Randbedingungen sie jeweils von Relevanz sind bzw. welche Faktoren auf die Güte der Modellierung Einfluss haben.

Im vorliegenden Beitrag soll es darum gehen, auf einen solchen Einflussfaktor aufmerksam zu machen, der im Grunde genommen nicht in der spezifischen Beschaffenheit des untersuchten Sprachmaterials begründet ist, sondern sozusagen bei dessen sekundärer Bearbeitung, nämlich der Aufbereitung für die theoretische Modellierung, ins Spiel kommt. Hierbei handelt es sich um eine in den entsprechenden Untersuchungen bislang in ihrer Auswirkung nicht hinreichend systematisch reflektierte Problematik, nämlich die Zusammenfassung von Satzlängenklassen zu bestimmten Intervallen.

1.2. Der Faktor der Intervallgröße als Einflussvariable der theoretischen Modellierung

Die Zusammenfassung von Satzlängen zu Intervallen ist ein übliches Verfahren, da die Messung der Satzlänge in der Anzahl der Worte pro Satz eine erhebliche Spannweite aufweist und die Satzlänge prinzipiell keiner quantitativen Beschränkung nach oben hin unterliegt. In der Satzlängenforschung hat sich daher das Verfahren eingebürgert, die Satzlängen in 5er-Intervalle von (*1-5, 6-10, ...*Wörter pro Satz) zusammenzufassen. So fasst bereits Yule (1939) in seiner Untersuchung zu Fragen des Stils und des Nachweises der Autorschaft die Satzlängen in 5er-Intervalle Wörtern zusammen (vgl. Yule 1939: 367). Auch Altmann (1988b) greift in seiner empirischen Untersuchung der von ihm postulierten Hyperpascal-Verteilung an 236 Texten die Satzlängen in 5er-Intervallen zusammen.

Vor dem Hintergrund dieses ehemals als selbstverständlich notwendig und unproblematisch angesehenen Vorgehens erweisen sich in einer Reihe jüngerer Untersuchungen zur (in der Anzahl der Worte pro Satz gemessenen) Satzlänge diese Zusammenfassungen (Intervallbildungen) als überaus problematisch:

1. Aus der Untersuchungen von Niehaus (2001: 210) zur Verteilung der Satzlänge an 20 deutschen literarischen Texten geht hervor, dass die von Altmann (1988a) theoretisch postulierte Hyperpascal-Verteilung nur dann passt, wenn man die Satzlängenklassen zu 5er-Intervallen zusammenfasst; ohne Zusammenfassung zu Intervallen erweist sich hingegen ein anderes Modell als adäquat, nämlich die negative Binomialverteilung
2. Einen ähnlichen Befund erhält Best (2001) als Ergebnis seiner Analyse der Satzlängenverteilung von 25 Texten der deutschen Gegenwartssprache. Seiner Interpretation nach stellen sich die Ergebnisse seiner Untersuchung folgendermaßen dar: „Für Satzlängen, gemessen nach der Zahl ihrer indirekten Konstituenten (Wörter), scheint die negative Binomialverteilung das beste Modell zu sein“. In dieser Studie von Best erwies sich die von Altmann (1988a) postulierte Hyperpascal-Verteilung allerdings als gänzlich ungeeignet, insofern der von Best beobachtete Befund unabhängig davon gilt, „ob man sie zu Fünfergruppen zusammenfasst oder nicht“ (Best 2001: 198).
3. In seiner Untersuchung der Satzlänge von Sprichwörtern in einem slowenischen Sprichwortkorpus testete Grzybek (1999) die Auswirkung unterschiedlicher Intervallbildungen (*1-2, 1-3, 1-4, ...*). Im Ergebnis zeigte sich, dass die Art der Zusammenfassung (d.h. die Größe der gebildeten Intervalle) nicht unbedingt einen Einfluss auf die theoretische Modellierung haben muss: Die Satzlängenverteilung folgte nämlich in allen Arten der Zusammenfassung der Hyperpoisson-Verteilung; nur für vollkommen

ungruppierte Satzlängen konnten keine guten Ergebnisse erzielt werden (vgl. Grzybek 1999: 104). Es muss jedoch im Hinblick auf diese Untersuchung angemerkt werden, dass es sich bei einer Sprichwortsammlung um spezifisches Datenmaterial handelt, welches aufgrund seiner einem Satz-Lexikon ähnlichen Struktur vermutlich anderen Gesetzmäßigkeiten der Satzlängenverteilungen unterliegt als etwa ein üblicher Fließtext.

4. In einer Folgestudie zur Verteilung der Satzlänge in einem Korpus deutscher Sprichwörter von Grzybek/Schlatte (2002) zeigte sich hingegen, dass die konkrete Art der Zusammenfassung einen eindeutigen Einflussfaktor hinsichtlich der theoretischen Modellierung darstellt – es wirkt sich nämlich „der konkrete Umfang der zusammengefassten Klassengrößen vehement auf das anzupassende Verteilungsmodell aus" (Grzybek/Schlatte 2002: 301). Jedoch sind interessanterweise weder die schon zuvor diskutierte Hyperpascal-Verteilung noch die von Grzybek (1999) für slowenisches Sprichwortmaterial herangezogene Hyperpoisson-Verteilung geeignet, die Satzlängen im Sprichwortkorpus ohne glättende Zusammenfassung zu modellieren. Erst bei einer Zusammenfassung zu 2er-Intervallen sind die Anpassungsergebnisse für beide Modelle akzeptabel; bei einer noch weiteren Zusammenfassung der Satzlängen zu 3er-, 4er-, und 5er-Klassen ist gar nur mehr die Hyperpoisson-Verteilung in Betracht zu ziehen.

Aus den angeführten Untersuchungen zur Satzlängenverteilung wird somit deutlich, dass die Zusammenfassung von Satzlängen in Intervalle offensichtlich als ein nicht zu vernachlässigender Einflussfaktor auf die theoretische Modellierung zu berücksichtigen ist: Einerseits zeigt es sich wiederholt, dass die ursprünglich von Altmann (1988a) postulierte Hyperpascal-Verteilung nur bei einer Zusammenfassung der Satzlängen in 5er-Intervalle als adäquates Modell der Satzlängenverteilungen in Betracht gezogen werden kann.[1] Aus den genannten empirischen Untersuchungen ergibt sich auch, dass selbst bei einer Zusammenfassung von (in der Anzahl der Worte pro Satz gemessenen) Satzlängen solche Modelle in Betracht zu ziehen sind, die für die Modellierung der Teilsatz/Satz-Ebene postuliert wurden (negative Binomial-Verteilung). Darüber hinaus wird die Hyperpoisson-Verteilung diskutiert, die sich in einschlägigen empirischen Untersuchungen wiederholt als ein adäquates Modell erwiesen hat.

Es leitet sich daher die Notwendigkeit ab, die Art und Weise der Zusammenfassung zu Satzlängenintervallen als einen eigenen Einflussfaktor einer systematischen Untersuchung zu unterziehen. Um diesen Zusammenhang zwischen der Zusammenfassung von Satzlängen zu Intervallen und der Adäquatheit theoretischer Modelle näher beleuchten zu können, wird im folgenden der von Grzybek (1999) vorgeschlagene empirische Ansatz zur Lösung dieses Problems aufgegriffen: Dieser besteht darin, dass zunächst an die beobachtete Verteilung der Satzlängen ohne jegliche Zusammenfassung zu Intervallen die in Frage kommenden Modelle angepasst werden; daran anschließend werden die Satzlängen aber auch in unterschiedlichen (1er-, 2er-, 3er-, 4er-, 5er-Intervallen) Intervallen zusammengefasst, bevor auch an diese geglätteten Daten die Modelle angepasst werden.

2. Pilotstudie zum Einflussfaktor der Intervallbildung am Beispiel slowenischer Texte

[1] In einer weiteren jüngeren Arbeit zur (in der Anzahl der Worte pro Satz gemessenen) Satzlänge (vgl. Kaßel/Livesey 2001) in englischen Texten werden die Satzlängenverteilungen ohne und mit Zusammenfassungen in 5er-Intervallen durchgeführt. Interessanterweise eignet sich in dieser Arbeit insbesondere die (1-verschobene) negative Binomialverteilung und die gemischte Poissonverteilung zur Modellierung der Satzlängenverteilungen; an gegebener Stelle wird jedoch die Hyperpascal-Verteilung – aus welchen Gründen auch immer – nicht diskutiert.

2.1. Modellierung der Satzlängen im Slowenischen

Die Verteilung der Satzlängen und die Frage der Beschreibung von Satzlängenverteilung durch theoretische Modelle ist im Slowenischen bislang wenig erforscht. In der Untersuchung von Jakopin (1999) gibt es einige wenige Hinweise auf die prozentuale Häufigkeitsverteilung von Satzlängen in slowenischen Texten. Beispielsweise wird gezeigt, dass die am häufigsten auftretende Satzlänge (prozentualer Anteil > 7%) in einem Korpus slowenischer literarischer Texte im Bereich von 4,5 – 6 Wörtern pro Satz liegt (vgl. Jakopin 1999). Des weiteren werden auch Angaben zu den längsten im Korpus auftretenden Sätzen gemacht. Die Ergebnisse sind jedoch für die vorliegende Untersuchung nicht von unmittelbarer Bedeutung, beziehen sich doch die Angaben der Satzlängen auf das jeweilige Gesamtkorpus und aufgrund der fehlenden Rohdaten ist auch keine Modellierung der Verteilung der Satzlängen möglich.

Die offensichtlich bislang einzige Untersuchung zur Verteilung der Satzlängen in slowenischen Texten stammt von Grzybek (1999). In dieser Studien wurde die Hypothese einer gesetzmäßigen Verteilung der (in der Anzahl der Worten pro Satz gemessenen) Satzlängen auf der Basis einer slowenischen Sprichwortsammlung überprüft. An diesem Material erwies sich die *Hyperpoisson-Verteilung* als besonders geeignet, um die Satzlängen in adäquater Weise zu modellieren (vgl. Grzybek 1999: 100). Zwar gilt zu beachten, dass eine Sammlung von Sprichwörtern vermutlich anderen Gesetzmäßigkeiten der Satzlängenverteilung folgt, als die hier vorliegenden literarischen Prosatexte, dennoch aber sollte auch diese Verteilung nicht außer acht bleiben. Dass sie in unmittelbarer Beziehung zu den oben bereits dargestellten negativen Binomialverteilung und zur Hyperpascal-Verteilung steht, lässt sich recht leicht zeigen: In Analogie zu den Ansätzen (4) und (5) stellt sich die Rekursionsformel der Hyperpoisson-Verteilung wie in (8) dar:

$$(8) \qquad P_x = \frac{b}{x+d} P_{x-1},$$

was in Analogie zu den Rekursionsformeln (6) und (7) in der folgenden Verteilung resultiert:

$$(9) \qquad P_x = \frac{b^x}{d^{(x)}{}_1F_1(1;d;b)}, \quad x = 0,1,2,...$$

Alle drei Verteilungsmodelle gehören insofern zu ein und derselben Familie, die man auch als erweiterte Katz-Familie bezeichnet: Es zeigt sich nämlich (vgl. Wimmer/Altmann 2000: 279ff., 449ff.), dass die negative Binomialverteilung ein Spezialfall der Hyperpascal-Verteilung ist (für $m = 1$), und dass die Hyperpascal-Verteilung gegen die Hyperpoisson-Verteilung konvergiert ($k \to \infty$, $q \to 0$, $kq \to b$).

Die empirische Überprüfung der Satzlängenverteilung mit Hilfe des Altmann-Fitters (2000) zeigte sich in der Tat, dass die folgenden theoretische Modelle in Erwägung zu ziehen sind:[2]

(a) die von Altmann (1988b) für die Satz/Wort-Variante postulierte *Hyperpascal-Verteilung*;

[2] In einer einleitenden Orientierungsphase wurde mit dem Altmann-Fitter (2000) zunächst exploratorisch (d.h. ohne theoretisch geleitete Vorannahmen) untersucht, welche Verteilungen überhaupt als geeignete Modelle in Frage kommen. Nach der Reduktion auf die drei oben genannten Verteilungen wurden die Detailanalysen auf diese beschränkt.

(b) die von Grzybek (1999) diskutierte *Hyperpoisson-Verteilung*;
(c) die von Best (2001) und Niehaus (2001) nachgewiesene *negative Binomialverteilung*.[3]

Im folgenden wird das Hauptaugenmerk auf den potentiellen Einflussfaktor der Zusammenfassung von Satzlängen zu Intervallen gerichtet; ungeachtet dieser Fokussierung lässt sich diese Frage nicht unabhängig von einer differenzierten Betrachtung der Analyseebenen (Korpus, komplexe Texte, Einzeltexte) verfolgen.

2.2. Empirische Untersuchung des Faktors Intervallgröße in slowenischen Texten

Als Basis für die empirische Untersuchung zum Faktor der Intervallgröße bei der Modellierung der Häufigkeitsverteilung von Satzlängen dient ein Korpus slowenischer Prosatexte. Diese setzte sich aus sechs Kurzromanen bzw. Kurzgeschichten (slowenisch: povest, im folgenden als Kurzerzählungen bezeichnet) von vier verschiedenen slowenischen Autoren aus dem 19. Jhd. zusammen (vgl. Tab. 1).

Tabelle 1
Slowenisches Textkorpus

Text	Autor	Titel
1	J. Kersnik	Mačkova očeta
2	J. Kersnik	Ponkrčev oča
3	I. Cankar	Hlapec Jernej in njegova pravica
4	J. Jurčic	Nemški Valpet
5	F. Levstik	Pokljuk
6	F. Levstik	Martin Krpan

Diese Texte werden im folgenden mit den entsprechenden Nummern (Text #1 bis Text #6) bezeichnet und als solche analysiert. Zum Zwecke der Qualitätskontrolle der Datenbasis – (zur Frage der Homogenität in quantitativen Untersuchungen vgl. Altmann 1992 bzw. Orlov 1982) – werden diese Texte jedoch nicht nur jeweils einzeln als komplexe Gesamttexte analysiert, sondern auf zwei weitere Arten und Weisen: Zum einen werden die genannten sechs Texte zu einem Gesamtkorpus zusammengefügt, so dass sich eine umfangreichere Textmischung ergibt; dieses Gesamtkorpus soll bedingt als „Text #7“ bezeichnet werden (vgl. Tab. 2). Zum anderen ergibt sich aufgrund der Tatsache, dass die Texte #2 und #3 jeweils aus mehreren Kapiteln bestehen, die Option, diese Kapitel als Einzeltexte zu verstehen und getrennt zu analysieren; in diesem Fall haben wir es mit den Texten #8 bis #28 (vgl. Tab. 3) zu tun.

Um die der im vorliegenden Text im Vordergrund stehende Problematik anschaulich zu illustrieren, finden sich im folgenden auf der Basis des gesamten Textkorpus die graphische Darstellungen 1a-1e: Während Abb. 1a die Daten ohne Zusammenfassung zu Intervallen ver-

[3] Interessanterweise hatte bereits Fucks (1970) in seinen Untersuchungen zur Satzlänge im Deutschen darauf hingewiesen, dass die negative Binomialverteilung sich gut eigne, die Verteilung von Satzlängen zu erfassen, wobei er allerdings die Satzlänge einer Reihe von deutschen Texten in der Anzahl der Silben pro Satz berechnet hatte. Dabei war Fucks allerdings von ganz anderen Voraussetzungen ausgegangen und hatte die Verteilung auf ganz andere Art und Weise abgeleitet: Ausgegangen war er nämlich von der Poisson-Verteilung, die dann die negative Binomialverteilung ergibt, wenn der Parameter der Poisson-Verteilung (der Mittelwert) einer Gammaverteilung folgt.

anschaulicht, stellen die Abb. 1b-1d das Ergebnis der unterschiedlichen Zusammenfassungen in den jeweiligen Intervallen dar, d.h. eine schrittweise Glättung.

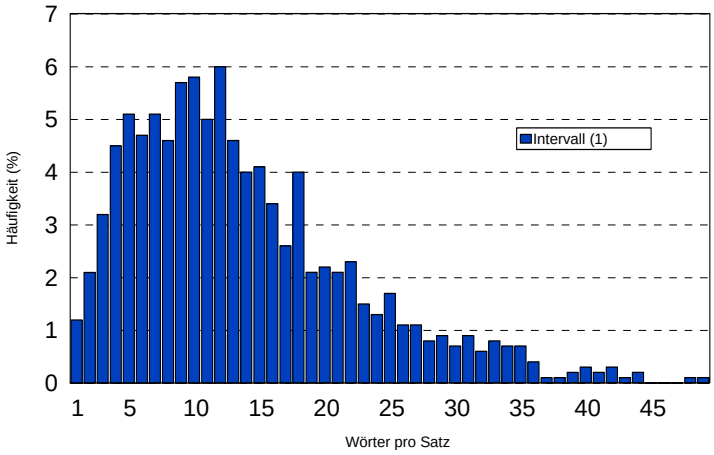

Abb. 1a: Ohne Zusammenfassung

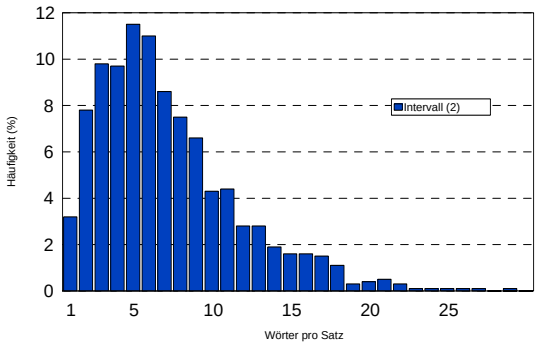

Abb. 1b: Zusammenfassung zu 2er Intervallen

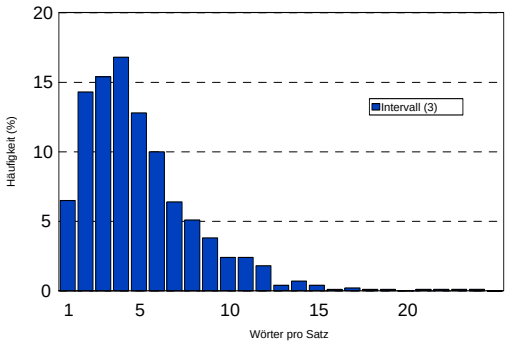

Abb. 1c: Zusammenfassung zu 3er Intervallen

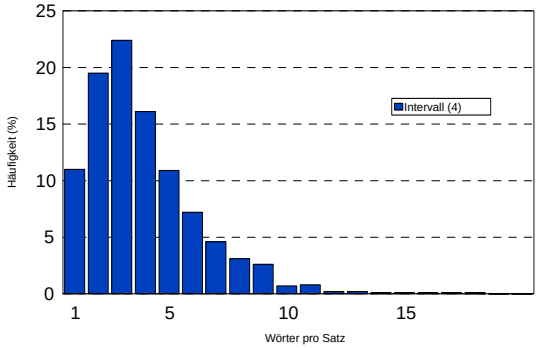

Abb. 1d: Zusammenfassung zu 4er Intervallen

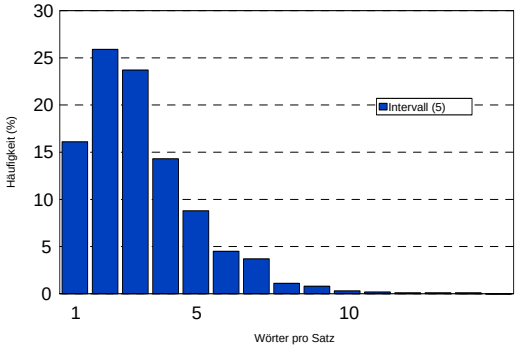

Abb. 1e: Zusammenfassung zu 5er Intervallen

Eine Zusammenfassung von Satzlängen zu Intervallen hat, wie auf den ersten Blick aus den graphischen Darstellungen ersichtlich wird, ganz offenbar folgende Auswirkungen:

a.) Einzelne Satzlängen, die sich durch eine besonders hohe Frequenz auszeichnen, machen sich mit zunehmender Intervallgröße nicht mehr so stark bemerkbar, wodurch insgesamt eine Nivellierung der Klassen bewirkt wird;

b.) Ebenso enthält mit zunehmender Intervallgröße die Häufigkeitsverteilungen immer weniger leere Satzlängenklassen (0 Worte pro Satz/Satzlängenintervall).

Dieser Befund ist als Ausgangspunkt für die folgende empirische Untersuchung des Faktors der Zusammenfassung von Satzlängen bei der theoretischen Modellierung zu verstehen.

2.3. Modellierung der Satzlängenhäufigkeit

Wie oben dargestellt, räumt die Spezifik der Textbasis die Möglichkeit ein, die Frage der Modellierung von Satzlängenhäufigkeiten unter Berücksichtigung des Einflussfaktors der Zusammenfassung zu Intervallen auf drei unterschiedlichen Analyseebenen zu untersuchen:

(1) Auf der ersten Ebene werden die genannten Kurzerzählungen zu einem Korpus zusammengefasst. Das Korpus, welches in Hinsicht auf die involvierten Textsorten als homogen zu bezeichnen ist, gibt die Möglichkeit zu prüfen, inwiefern eine Korpusanalyse gegebenenfalls anderen Gesetzmäßigkeiten folgt als die Analyse der einzelnen Texte. Insgesamt besteht das Korpus aus 2758 Sätzen mit 38966 Wörtern; die einzelnen Werte sind in der Tab. 2 zusammengefasst.[4]

Tabelle 2
Quantitative Angaben zum Textkorpus

Textnr.	**Sätze**	**Wörter**	**min.**	**max.**	$\square$
# 7	2758	39016	2	106	14.15

(2) Auf der zweiten Ebene werden die sechs Kurzerzählungen als komplexe Texte aufgefasst; anzumerken ist, dass dabei eine textinterne, von den Autoren selbst vorgenommene Kapitelgliederung nicht beachtet wird. Insgesamt handelt es sich also um sechs unterschiedliche Kurzerzählungen von vier verschiedenen Autoren, wobei Text #3 mit insgesamt 1383 Sätzen den größten Umfang aufweist. Unter dieser Voraussetzung zeichnen sich die Texte durch die in Tab. 3 zusammengefassten Charakteristika aus.

Tabelle 3
Quantitative Angaben zu den komplexen Texten

Text	**Autor**	**Titel**	**Sätze**	**Wörter**	**min.**	**max.**	$\bar{x}$
# 1	J. Kersnik	Mačkova očeta	109	1597	1	61	14.65
# 2	J. Kersnik	Ponkrčev oča	165	2178	1	42	13.20
# 3	I. Cankar	Hlapec Jernej in njegova pravica	1383	18407	1	97	13.31
# 4	J. Jurčic	Nemški Valpet	561	7921	1	63	14.21
# 5	F. Levstik	Pokljuk	169	3181	2	70	18.82
# 6	F. Levstik	Martin Krpan	371	5682	2	106	15.32

(3) Auf der dritten Ebene werden die Texte unter Berücksichtigung der von den Autoren selbst vorgenommen Kapitelgliederung jeweils individuell untersucht. Im Detail weist der Text von Janko Kersnik „Ponkrčev oča“ drei Kapitel auf; Ivan Cankars „Hlapec Jernej ...“

[4] Neben der jeweiligen Textnummer sowie dem Autor und Titel des Werks finden sich die Anzahl der Sätze, Minimum und Maximum der Wortanzahl pro Satz sowie die durchschnittliche Satzlänge.

besteht insgesamt aus 18 Einzelkapiteln (insgesamt also 21 Einzelkapitel). Tabelle 4 resümiert die wesentlichen Charakteristika der Einzeltexte.

Tabelle 4
Quantitative Angaben zu den Einzeltexten

Text	**Autor**		**Titel**	**Sätze**	**Wörter**	**min.**	**max.**	$\bar{x}$
# 8	J. Kersnik	Ponkrčev oča	Kapitel 1	68	895	1	40	13.16
# 9			Kapitel 2	38	523	4	35	13.76
# 10			Kapitel 3	59	760	1	42	12.88
# 11	I. Cankar	Hlapec Jernej	Kapitel 1	43	602	2	41	14.00
# 12			Kapitel 2	82	977	1	36	11.91
# 13			Kapitel 3	85	1038	1	50	12.21
# 14			Kapitel 4	57	796	3	97	13.96
# 15			Kapitel 5	80	809	1	37	10.11
# 16			Kapitel 6	75	890	2	36	11.87
# 17			Kapitel 7	71	973	3	41	13.70
# 18			Kapitel 8	107	1473	1	37	13.77
# 19			Kapitel 9	60	939	5	35	15.65
# 20			Kapitel 10	113	1134	1	39	10.04
# 21			Kapitel 11	75	937	1	33	12.49
# 22			Kapitel 12	80	1203	1	78	15.04
# 23			Kapitel 13	119	1583	1	51	13.30
# 24			Kapitel 14	53	956	2	72	18.04
# 25			Kapitel 15	98	1388	1	36	14.16
# 26			Kapitel 16	77	1203	1	66	15.62
# 27			Kapitel 17	87	1203	1	33	13.83
# 28			Kapitel 18	21	303	1	38	14.43

Durch diese Gliederung auf drei unterschiedliche Analyseebenen ergeben sich insgesamt 28 Datensätze, in denen die Satzlänge bestimmt werden kann. Auf Basis dieser Texte lässt sich nunmehr – neben der Frage des Einflussfaktors Intervallgröße – auch die Frage der Datenhomogenität untersuchen.

Exkurs:
Zur Definition des Satzes und der automatischen Berechnung der Satzlänge

Die Satzlänge in den genannten Texten wurde automatisiert analysiert und berechnet (zur genaueren Bestimmung siehe Kelih/Grzybek 2004). Ausgehend von der Möglichkeit der formalen Bestimmung der Satzgrenzen aufgrund von Interpunktionszeichen wird dem Punkt, dem Ausrufe- sowie dem Fragezeichen eine satzabgrenzende Funktion zugesprochen. In Anbetracht der Tatsache, dass der Punkt im Slowenischen jedoch in einigen Positionen keine satzabgrenzende Funktion hat (z.B. als Abkürzung wie in „*c.kr*.“ oder als Auslassungszeichen wie in “*To je Sitarjevo… tam!*”) und unter Berücksichtigung des Umstandes, dass Frage- und Ausrufezeichen auch zur Kennzeichnung von Interjektionen dienen können (vgl. Pravopis 1990: 38ff.), sind einige Modifizierungen notwendig.

In Anlehnung an die grundsätzlichen Überlegungen von Grinbaum (1996) zur Automatisierung von Satzlängenuntersuchungen bietet es sich daher an, den Großbuchstaben als weiteres satzabgrenzendes Zeichen in eine formal bestimmbare Satzdefinition einzubauen. Wenn auch der Großbuchstabe im Slowenischen zur Kennzeichnung von Eigennamen, geographischen Bezeichnungen und ähnlichem dient, liegt eine weitere zentrale Funktion des Großbuchstabens darin, den Anfang von Texten, Absätzen und einzelnen Sätzen zu markieren. In dieser Funktion als Gliederungsmerkmal von Texten können Großbuchstaben bei der Bestimmung von Satzgrenzen herangezogen werden (vgl. Grinbaum 1996: 454). Somit sind die Interpunktionszeichen (.), (…), (?) und (!) in Kombination mit einem Großbuchstaben am Anfang des nächstfolgenden Satzes eindeutig als satzabschließend zu identifizieren. Da jedoch den erwähnten Interpunktionszeichen nicht in allen Fällen ein Buchstabe folgen muss (Textende, Absatzende) gelangt man zu der folgenden Satzdefinition, die auch der vorliegenden Untersuchung zur Anwendung kommt:

> **Als Satzendezeichen gelten [.], […], [!] und [?], es sein denn, ein Kleinbuchstabe ist das erste Graphem des nächsten Wortes.**[5]

Aufgrund der verwendeten Satzdefinition ist es möglich, den Satz (und damit dann auch die Satzlänge) automatisiert zu bestimmen. Das vorliegende Textkorpus und die vorgestellte Satzdefinition sind nunmehr als Ausgangspunkt für die empirische Untersuchung der Satzlängenverteilung in slowenischen Texten zu sehen.

2.4. Einflussfaktor Satzlängenintervalle bei der theoretischen Modellierung

Unter den dargestellten Voraussetzungen wird im nächsten Schritt somit zu prüfen sein, ob für die erwähnten theoretischen einzelnen Verteilungsmodelle die vorgenommene Zusammenfassung von Satzlängen in Intervalle einen Einflussfaktor darstellt. Entsprechend der obigen Ausführungen sollen nunmehr die Ergebnisse der theoretischen Modellierung – getrennt für die unterschiedenen Analyseebenen – präsentiert werden. Als Güte der Anpassung wird dabei die Überschreitungswahrscheinlichkeit P des χ^2-*Tests* angeführt: Dabei wird ein Wert von $P \geq 0.01$ als Kennwert einer akzeptablen Übereinstimmung von empirischen Daten und theoretischem Modell angesehen; für längere Texte wird der Kontingenzkoeffizient $C = \chi^2 / N$ (vgl. Grotjahn/Altmann 1993) in Betracht gezogen, wobei $C \leq 0.02$ als gute, $C \leq 0.01$ als sehr gute Anpassung des jeweiligen theoretischen Modells betrachtet wird (die jeweiligen P- und C-Werte vgl. Anhang 1-2)[6].

---

[5] Natürlich kommen bei derartigen Definition sprachspezifische bzw. kulturspezifisch-typographische Konventionen ins Spiel. So wird in der Arbeit auf russische Texte bezogenen Arbeit von Kelih/Grzybek (2004) der Satz alternativ definiert als „eine durch Punkt, Frage- und Ausrufezeichen abgegrenzte Einheit des Textes", wobei diese einfachere Definition und die hier angeführten einem statistischen Vergleich unterzogen werden. In der genannten Arbeit kann gezeigt werden, dass die Anwendung von unterschiedlichen Satzdefinitionen zu keinen statistisch signifikanten Unterschieden führt. Auch auf der Ebene der theoretischen Modellierung von Satzlängenverteilungen spielen die beiden unterschiedlichen Satzdefinitionen keine signifikante Rolle.

[6] Die gesamten Rohdaten zur vorliegenden Untersuchung finden sich in Kelih (2002).

2.4.1. Theoretische Modellierung auf Korpusebene

Auf der Ebene des Gesamtkorpus (vgl. Tab. 2) ergibt sich erstes Ergebnis, dass die Hyperpoisson-Verteilung auf der Ebene des Gesamtkorpuses ein gänzlich unpassendes Modell darstellt. Im Gegensatz dazu weist die negative Binomialverteilung – außer bei jeglicher Nicht-Zusammenfassung von Satzlängen zu Intervallen – eine gute Übereinstimmung mit dem Modell auf (vgl. Tabelle 5 mit den *C*-Werten). Das insgesamt überzeugendste Ergebnis liefert bei Nicht-Zusammenfassung ebenso wie bei allen Arten der Intervallbildung die Hyperpascal-Verteilung, die ja von Altmann (1988a) für die (in der Anzahl der Worten pro Satz berechneten) Satzlängenverteilung postuliert worden war.

Tabelle 5
C-Werte für das Gesamtkorpus

Verteilung	**Intervall**				
	1	2	3	4	5
Negative Binomial	0.0982	0.0079	0.0102	0.0083	0.0048
Hyper-Poisson	0.0802	0.0621	0.051	0.0375	0.0262
Hyper-Pascal	0.0161	0.0130	0.0164	0.0151	0.0133

2.4.2. Theoretische Modellierung auf Ebene der komplexen Texte

In einem zweiten Schritt folgt die Analyse der sechs komplexen Texte. Auf dieser Ebene kann bereits die Frage gestellt werden, ob die Satzlängenverteilungen bzw. deren zugrunde liegenden Modelle denen des Korpus entsprechen, oder ob es hier zu Unterschieden in der Eignung der theoretischen Modelle kommt. Aufgrund der umfangreichen Daten, werden die jeweiligen *P*-Werte bzw. (für Text # 3) der *C*-Werte im Anhang (Tabelle 1.1 und 1.2) zusammengefasst.

An dieser Stelle soll lediglich eine tabellarische Übersicht des absoluten und prozentualen Anteils der Texte, die eine zufriedenstellende Übereinstimmung mit dem Modell zeigen, geboten werden.

Tabelle 6
Anteil von Texten mit $P \geq 0.01$ bzw. $C \leq 0.02$ (abs. und prozentual)

	Intervalle									
	1		2		3		4		5	
Verteilung	abs.	%	abs.	%	abs	%	abs.	%	abs	%
Negativ-Binomial	6	100	6	100	5	83	5	83	4	67
Hyper-Poisson	5	83	5	83	4	67	4	67	4	67
Hyper-Pascal	1	17	2	33	4	67	6	100	5	83

Die Analyse der theoretischen Modelle für die Satzlängen der sechs abgeschlossenen slowenischen Kurzgeschichten zeigt folgende Ergebnisse: Die Häufigkeit der Satzlängen kann in dieser Textauswahl nicht nur durch ein einziges Modell beschrieben werden, da die Ergebnisse stark in Abhängigkeit von den vorgenommenen Zusammenfassungen variieren. So ist die *negative Binomialverteilung* für diese Texte nur ohne Zusammenfassungen bzw. bei einer Zusammenfassung zu 2er-Intervallen geeignet; eine weitere Zusammenfassung bewirkt jedoch eine Verschlechterung der Ergebnisse. Die *Hyperpoisson-Verteilung* liefert auch auf dieser Ebene keine überzeugenden Ergebnisse. Für die *Hyperpascal-Verteilung* schließlich ergibt sich der interessante Befund, dass mit zunehmender Zusammenfassung der Anteil von Texten, die dieser Verteilung folgen, steigt (außer bei Zusammenfassungen zu 5er-Intervallen). Damit zeichnet sich insgesamt an dieser Stelle der Trend ab, dass sowohl die negative Binomialverteilung als auch die Hyperpascal-Verteilung als geeignete Modelle anzusehen sind. Zu beachten ist jedoch, dass die Zusammenfassung von Satzlängen zu Intervallen auf dieser Analyseebene einen Einfluss auf die Adäquatheit bestimmter theoretischer Modelle hat.

2.4.3. Theoretische Modellierung auf Ebene von Einzeltexten

Im dritten und letzten Schritt sollen nunmehr die jeweiligen im obigen Sinne definierten Einzeltexte (vgl. Tab. 4) analysiert werden. Es handelt sich hierbei um insgesamt 21 einzelne Texte, die sich aus den einzelnen Kurzgeschichten ergeben. Wie in Tab. 7 dargestellt, zeigt sich auf dieser Analyseebene ein relativ klares Bild. So gilt es als erstes allgemein festzustellen, dass die Anpassungsergebnisse für die einzelnen Texte insgesamt besser sind als für die komplexen Texte – dies ist ein starkes Argument für die Homogenität der Daten auf dieser Analyseebene, und ein Grund für die Annahme, dass wir es nicht nur auf der Ebene des Gesamtkorpus, sondern auch schon auf der Ebene der komplexen Texte mit in unterschiedlichen Regimes resultierenden Text-Mischungen zu tun haben könnten. Im Detail kann dann weiterhin festgehalten werden, dass sowohl die *negative Binomialverteilung* als auch die *Hyperpoisson-Verteilung* bei allen Arten der Zusammenfassung passende Modelle für die Häufigkeitsverteilung der Satzlängen sind.

Tabelle 7
Anteil von Kapiteln mit $P \geq 0.01$ (abs. und prozentual)

	Intervall									
	1		2		3		4		5	
Verteilung	abs.	%	abs.	%	abs	%	abs.	%	abs	%
Negativ-Binomial	21	100	21	100	21	100	21	100	21	100
Hyper-Poisson	21	100	21	100	21	100	21	100	21	100
Hyper-Pascal	2	10	11	52	14	67	14	67	16	76

Für die *Hyperpascal-Verteilung* ist festzustellen, dass dieser ohne Zusammenfassung nicht mehr als 10% der Texte folgen; erst eine Zusammenfassung zu Intervallen ergibt eine Verbesserung des Ergebnisses, wobei allerdings auch im 5er-Intervall nicht mehr als 76% der Texte dieser Verteilung folgen. Insgesamt zeigt es sich also, dass auf dieser Ebene die Hyper-

pascal-Verteilung eine erhebliche „Sensibilität“ gegenüber der Art der Zusammenfassung aufweist. Zusammenfassend ist also davon auszugehen, dass die Art der Zusammenfassung als ein nicht unerheblicher Einflussfaktor der Satzlängenmodellierung in Betracht zu ziehen ist; es zeigt sich, dass insbesondere die Hyperpascal-Verteilung davon betroffen ist.

2.4.4. Gesamtauswertung der Ergebnisse

Im Anschluss an die Durchführung der Untersuchungen auf den drei unterschiedlichen Ebenen, bei der sich bereits deutliche Tendenzen abgezeichnet haben, können nunmehr die Ergebnisse im Hinblick auf das Gesamtkorpus, die sechs Kurzgeschichten und die 21 Einzelkapitel auch zusammenfassend dargestellt werden – dies freilich nur im Sinne einer allgemeinen Orientierung, da auch weiterhin davon auszugehen ist, dass Korpusanalysen und Einzeltextanalysen nicht zu ein und denselben Ergebnissen führen (müssen). Wie der Tab. 8 zu entnehmen ist, führt die Anpassung der *negativen Binomialverteilung* insgesamt zu den besten Ergebnissen. Dies gilt grosso modo auch für die *Hyperpoisson-Verteilung*, die sich auf der Ebene des Gesamtkorpus als gänzlich unpassend erwies, aber auf der Ebene der abgeschlossenen Kurzgeschichten und der Einzelkapitel als Modell der Satzlängenverteilung als Modell in Frage kommt. Für die *Hyperpascal-Verteilung* bestätigt sich insgesamt der bereits beobachtete Trend, dass die Zusammenfassung von (in der Anzahl der Worte pro Satz gemessenen) Satzlängen zu Intervallen einen entscheidenden Einflussfaktor darstellt: Erst durch die Zusammenfassung von Satzlängen erweist sich die Hyperpascal-Verteilung als ein passendes Modell, wobei relativierend anzumerken ist, dass die Anpassungsgüte der Hyperpascal-Verteilung in keinem Fall die guten Ergebnisse der beiden anderen Modelle übertreffen kann.

Tabelle 8
Zusammenfassung der Ergebnisse
($P \geq 0.01$ bzw. $C \leq 0.02$; absolute und prozentualer Anteil)

	Intervall									
	1		2		3		4		5	
Verteilung	abs.	%	abs.	%	abs	%	abs.	%	abs	%
Negative Binomial	27	96	28	100	27	96	27	96	26	93
Hyper-Poisson	26	93	26	93	25	89	25	89	25	89
Hyper-Pascal	4	14	14	50	19	68	21	75	22	79

Das Gesamtergebnis lässt sich wie in Abb. 1 anschaulich darstellen.

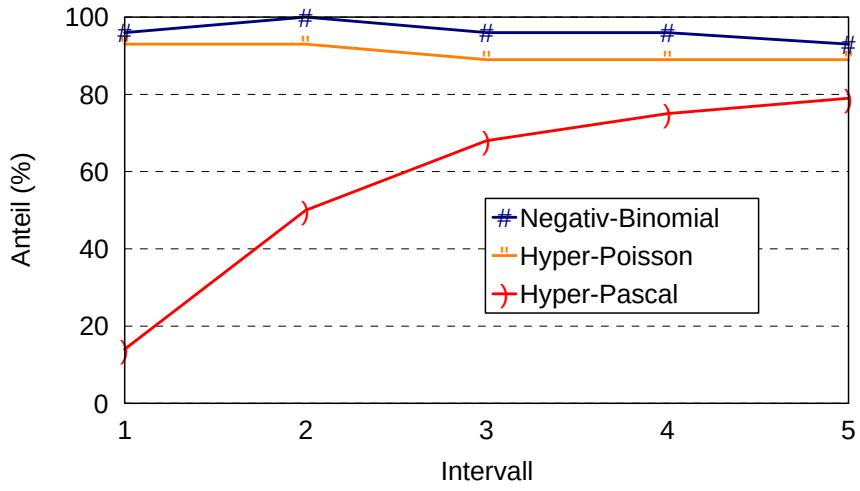

Abb. 1. Graphische Darstellung der Ergebnisse auf dem Signifikanzniveau $\alpha = 0.01$ ($P > 0.01$ bzw. $C \leq 0.02$; relativer Anteil)

3. Resümee

Die vorliegenden Untersuchungen zur Satzlänge bestätigen ein weiteres mal die hypothetisch formulierte Gesetzmäßigkeit der Häufigkeitsverteilung von Satzlängen (vgl. Altmann 1988, 1988b), gemessen in der Anzahl der Worte pro Satz. Drei in der Vergangenheit postulierte und in empirischen Arbeiten wiederholt als relevant nachgewiesene Modelle (Hyperpascal-Verteilung, Hyperpoisson-Verteilung und negative Binomialverteilung) wurden an insgesamt 28 slowenischen Datensätzen überprüft. Durch die systematische Überprüfung der Satzlängenverteilung auf drei unterschiedlichen Analyseebenen (Korpus, komplexe Texte in Form von Kurzerzählungen, sowie Einzeltexte auf der Ebene einzelner Kapitel) konnte gezeigt werden, dass von der Gattung her relativ homogene Texte unterschiedlicher Autoren im Hinblick auf die Häufigkeitsverteilung der Satzlängen durch ein einheitliches Modell beschrieben werden können.

1. Im Gesamtergebnis zeigt die *negative Binomialverteilung* insgesamt die „besten" Ergebnisse, insofern sie sich auf allen drei Analyseebenen bestens zur theoretischen Modellierung eignet und dies unabhängig von der Frage, ob und wie die Daten zu Intervallen zusammengefasst werden. Interessanterweise handelt es sich hierbei um genau jenes Modell, das Altmann (1988b) für die Verteilung der Satzlängen postuliert hat, wenn diese in der Anzahl der Teilsätze (und nicht, wie hier, in der Anzahl der Worte) pro Satz gemessen werden. Es ist daher anzunehmen, dass die Wortebene nicht prinzipiell als zusätzlicher Störfaktor bei der Modellierung der Satzlängen zu betrachten ist. Insofern bestätigt sich die auch an deutschem Sprachmaterial vorgenommene Interpretation „dass die intervenierende Ebene bei der Satz-Wort-Variante im Deutschen offenbar keine Störungen hervorruft, wie dies von Altmann noch generell vermutet wurde" (Niehaus 2001: 211).
2. Als in Frage kommendes Modell unbedingt in Betracht zu ziehen ist – zumindest (!) für slowenische Texte – auch die *Hyperpoisson-Verteilung*, die sich im Grunde genommen bei allen komplexen und individuellen Texten als bestens geeignet erwiesen hat und als Modell lediglich bei der Analyse des Gesamtkorpus nicht in Frage kam. Hier stellt sich heraus, dass ganz offenbar die Analyseebene selbst – zumindest auf der

Basis des hier untersuchten Datenmaterials – einen entscheidenden Einfluss auf die theoretische Modellierung von Satzlängenverteilungen haben kann.

3. Für die *Hyperpascal-Verteilung* schließlich konnte auf empirischen Wege gezeigt werden, dass die Zusammenfassung zu Satzlängenintervallen als ein wesentlicher Einflussfaktor bei der theoretischen Modellierung von Satzlängenverteilungen anzusehen ist. Hier gilt, dass sie unabhängig von der Analyseebene dann (und nur dann) als geeignetes Modell ins Spiel kommt, wenn die Daten zum Zwecke der Glättung zu größeren (4er- oder 5er-) Intervallen zusammengefasst werden.

Diese Schlussfolgerungen gelten zunächst einmal nur für das von uns analysierte slowenische Material.[7] Die Tragweite dieser Schlussfolgerung wird in Zukunft nicht nur an umfangreicherem Material unter Berücksichtigung auch anderer Textsorten zu überprüfen sein, sondern auch an Material aus anderen Sprachen. Dennoch sollte mit den Ergebnissen der vorliegenden Studie nachhaltig nicht nur auf die Problematik der Analyseebene, sondern auch auf die Auswirkung datenglättender Zusammenfassungen in Form von Intervallbildungen aufmerksam gemacht worden sein.

Literatur

Altmann, G. (1988a). *Wiederholungen in Texten (= Quantitative Linguistics, Vol. 36).* Bochum: Brockmeyer.

Altmann, G. (1988b). Verteilungen der Satzlängen. In: K.P. Schulz (Hrsg.), *Glottometrika 9, 147-169.*

Altmann, G. (1992). Das Problem der Datenhomogenität. In: B. Rieger (Hrsg.), *Glottometrika 13, 287-298.*

Best, K.H. (2001). Wie viele Wörter enthalten Sätze im Deutschen? Ein Beitrag zu den Sherman-Altmann-Gesetzen. In: K.H. Best (Hg.), *Häufigkeitsverteilungen in Texten: 167-201.* Göttingen: Peust & Gutschmidt.

Fucks, W. (1970). Analyse formaler Eigenschaften von Texten mit mathematischen Hilfsmitteln. In: Borck, Karl Heinz; Henss, Rudolg (Hg.), *Der Berliner Germanistentag 1968. Vorträge und Berichte: 42-52*. Heidelberg: Winter.

Grinbaum, O.N. (1996). Komp'juternye aspekty stilemetrii. In: A.S. Gerd (Hrsg.), *Prikladnoe jazykoznanie: 451-463.* Sankt Peterburg: Izdatel'stvo S.-Peterburgskogo universiteta.

Grotjahn, R., Altmann, G. (1993). Modelling the Distribution of Word length: Some Methodological Problems. In: Köhler, R. and Rieger, B.B. (eds.), *Contributions to Quantitative Linguistics: 141-153.* Dordrecht: Kluwer.

Grzybek, P. (1999). Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter. *Anzeiger für Slavische Philologie XXVII, 87-108.*

Grzybek, P., Schlatte, R. (2002). Zur Satzlänge deutscher Sprichwörter. Ein Neuansatz. In: Piirainen, E., Piirainen, I.T. (Hrsg.), *Phraseologieforschung in Raum und Zeit: 273-284.* Baltmannsweiler: Schneider.

Jakopin, F. (1999). *Zgornja meja entropije pri leposlovnih besedelih v slovenskem jeziku.* Dokotorska disertacija. Ljubljana. [http://www.ff.uni-lj.si/hp/pj/disertacija/]

Kaßel, A., Livesey, E. (2001). Untersuchungen zur Satzlängenhäufigkeit im Englischen: Am Beispiel von Texten aus Presse und Literatur (Belletristik). *Glottometrics 1, 27-50.*

[7] Dass die Tendenzen auch bei russischen Texten ganz ähnlich sind wie hier in Bezug auf slowenische Texte dargestellt, wurde bei Kelih (2002) nachgewiesen.

Kelih, E., Grzybek, P. (2004). Satzlänge: Definition, Häufigkeiten, theoretische Modellierung. In: A. Mehler (Ed.), *Quantitative Methoden in Computerlinguistik und Sprachtechnologie.* [= Special Issue of: LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie // Journal for Computational Linguistics and Language Technology]

Kelih, E. (2002). *Untersuchungen zur Satzlänge in russischen und slowenischen Prosatexten.* Diplomarbeit in 2 Bänden. Graz.
[vgl. http://www-gewi.uni-graz.at/quanta/research/quanta_sentence.htm]

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

Niehaus, B. (2001). Die Satzlängenverteilung in literarischen Prosatexten der Gegenwart. In: L. Uhlířová, G. Wimmer, G. Altmann & R. Köhler (Eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček: 196-214.* Trier: WVT.

Orlov, Ju. K. (1982). Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache – Rede" in der statistischen Linguistik). In: Orlov, Ju.K., Boroda, M.G., Nadarševili, I.Š. (1982), *Sprache, Text, Kunst. Quantitative Analysen: 1-55*. Bochum: Brockmeyer.

Pravopis (1990). *Slovenski pravopis. Pravila*. Ljubljana: ZRC SAZU.

Sherman, L.A. (1888). Some observations upon the sentence-length in English prose. *The University of Nebraska Studies 1, 119-130.*

Sichel, H.S. (1974). On a distribution representing sentence length in written prose. *Journal of the Royal Statistical Society A 137, 25-34.*

Williams, C.B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika 31, 356-361.*

Wimmer, G., Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In: P. Schmidt, (ed.), *Issues in General Linguistic Theory and The Theory of Word Length: 112-133* [= Glottometrika 15.]. Trier: WVT.

Wimmer, G., Altmann, G. (2000). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

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.*

Anhang

1. Anpassungsergebnisse an die sechs komplexen Texte

1.1. *C*-Werte für Text # 3

Verteilung	**Intervalle**				
	1	2	3	4	5
Negativ- Binomial	0.0086	0.0146	0.0184	0.0165	0.0121
Hyper-Poisson	0.0034	0.009	0.0199	0.0114	0.0040
Hyper-Pascal	0.0698	0.0692	0.0488	0.0038	0.0197

1.2. *P*-Werte für die Texte #1, #2, #4, #5, #6

	Negativ Binomial				
	Intervalle				
Textnr.	**1**	**2**	**3**	**4**	**5**
# 1	0.3767	0.7246	0.6146	0.8146	0.8274
# 2	0.4680	0.0610	0.0569	0.1324	0.2942
# 4	0.1845	0.2230	0.0001	0.0000	0.0020
# 5	0.4646	0.7573	0.6611	0.5292	0.2364
# 6	0.0555	0.0543	0.0532	0.0576	0.0005

	Hyperpoisson				
	Intervalle				
Textnr.	**1**	**2**	**3**	**4**	**5**
# 1	0.3041	0.5704	0.3633	0.8273	0.6818
# 2	0.0570	0.0250	0.4072	0.0681	0.0597
# 4	0.0784	0.5728	0.0000	0.0000	0.0000
# 5	0.0737	0.1756	0.1394	0.1025	0.0783
# 6	0.0000	0.0000	0.0000	0.0001	0.0000

	Hyperpascal				
	Intervalle				
Textnr.	**1**	**2**	**3**	**4**	**5**
# 1	0.000	0.000	0.485	0.504	0.0000
# 2	0.000	0.000	0.089	0.324	0.4159
# 4	0.0117	0.000	0.126	0.038	0.0167
# 5	0.0000	0.66	0.456	0.537	0.1002
# 6	0.0000	0.055	0.0000	0.061	0.1088

3. Anpassungsergebnisse an die 21 Einzeltexte (*P*-Werte)

Text	Negativ Binomial					Hyperpoisson				
	Intervalle					Intervalle				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
8	0.7505	0.193	0.5232	0.384	0.38	0.8114	0.181	0.5278	0.427	0.418
9	0.2347	0.444	0.2512	0.152	0.088	0.1278	0.113	0.1373	0.076	0.078
10	0.9002	0.34	0.0647	0.407	0.729	0.1555	0.102	0.3331	0.173	0.433
11	0.1834	0.201	0.3359	0.524	0.061	0.1342	0.056	0.2656	0.103	0.087
12	0.9644	0.811	0.9951	0.643	0.971	0.899	0.607	0.9595	0.696	0.92
13	0.5113	0.19	0.0634	0.558	0.072	0.2876	0.101	0.0135	0.453	0.055
14	0.1085	0.117	0.2019	0.147	0.123	0.0397	0.055	0.1271	0.241	0.041
15	0.1846	0.96	0.4997	0.852	0.426	0.1943	0.909	0.4469	0.912	0.585
16	0.6772	0.445	0.3704	0.293	0.07	0.1421	0.109	0.0776	0.089	0.427
17	0.7047	0.957	0.816	0.831	0.96	0.4694	0.708	0.5163	0.705	0.85
18	0.1201	0.425	0.0942	0.557	0.432	0.3688	0.754	0.2705	0.551	0.383
19	0.6664	0.991	0.278	0.085	0.493	0.8542	0.156	0.2861	0.166	0.371
20	0.0518	0.144	0.1543	0.06	0.045	0.1546	0.152	0.3713	0.113	0.089
21	0.1321	0.2	0.5894	0.653	0.135	0.2337	0.633	0.7811	0.815	0.251
22	0.8937	0.968	0.8915	0.688	0.885	0.7866	0.894	0.8523	0.608	0.884
23	0.8972	0.985	0.979	0.991	0.958	0.7177	0.937	0.8683	0.975	0.977
24	0.8782	0.644	0.1896	0.197	0.208	0.7422	0.385	0.1095	0.079	0.196
25	0.6377	0.663	0.8175	0.84	0.837	0.8135	0.819	0.8733	0.836	0.923
26	0.7735	0.332	0.1256	0.094	0.066	0.2599	0.078	0.453	0.453	0.799
27	0.5982	0.216	0.4004	0.271	0.79	0.8269	0.359	0.5092	0.277	0.816
28	0.6333	0.457	0.1633	0.444	0.156	0.6314	0.669	0.2288	0.471	0.181

Text	**Hyperpascal**					**Text**	**Hyperpascal**				
	Intervalle						**Intervalle**				
	1	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
8	0.0000	0.0513	0.3923	0.2690	0.0000	19	0.0000	0.0000	0.0587	0.0673	0.0688
9	0.0000	0.0117	0.4397	0.2097	0.3628	20	0.1598	0.0122	0.0477	0.0056	0.0126
10	0.0000	0.0815	0.1128	0.6353	0.8105	21	0.0000	0.0000	0.0091	0.0000	0.0002
11	0.1128	0.0000	0.0000	0.0471	0.1996	22	0.0000	0.7619	0.0413	0.1298	0.9668
12	0.0000	0.7570	0.9834	0.0000	0.9366	23	0.0000	0.4627	0.8041	0.9696	0.0000
13	0.0000	0.0531	0.0211	0.2871	0.0025	24	0.0000	0.0000	0.0000	0.0000	0.0000
14	0.0000	0.0000	0.0000	0.2351	0.5980	25	0.0000	0.0000	0.2062	0.7074	0.6542
15	0.0000	0.9266	0.2197	0.5254	0.4243	26	0.0000	0.0000	0.0000	0.1198	0.1392
16	0.0000	0.7455	0.4821	0.3198	0.3881	27	0.0000	0.0010	0.0000	0.0000	0.5115
17	0.0000	0.9890	0.8100	0.6245	0.8248	28	0.0000	0.0000	0.0830	0.0000	0.0712
18	0.0000	0.0000	0.0000	0.0000	0.2518						

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Herausgeber – Editors

G. Altmann	02351973070-0001@t-online.de
K.-H. Best	kbest@gwdg.de
P. Grzybek	peter.grzybek@uni-graz.at
A. Hardie	a.hardie@lancaster.ac.uk
L. Hřebíček	hrebicek@orient.cas.cz
R. Köhler	koehler@uni-trier.de
V. Kromer	kromer@newmail.ru
O. Rottmann	otto.rottmann@t-online.de
A. Schulz	reuter.schulz@t-online.de
G. Wimmer	wimmer@mat.savba.sk
A. Ziegler	arneziegler@compuserve.de

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –8 (2004) –. – Lüdenscheid: RAM-Verl., 2004
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 8 (2004)
ISSN 1617-8351

Contents

History of quantitative linguistics

Available issues