Häufigkeiten von Wortlängen und Wortlängenpaaren: Untersuchungen am Beispiel russischer Texte von Viktor Pelevin

PETER GRZYBEK / EMMERICH KELIH (Graz)

Theoretische Vorüberlegungen

Die Länge eines Wortes und die Häufigkeit, mit der Wörter einer bestimmten Länge in Texten oder Textmengen vorkommen, ist seit mehr als 100 Jahre als eines der Merkmale angesehen worden, mit dem man spezifische stilistische Eigenschaften von Autoren, von individuellen Texten oder aber auch von bestimmten Texttypen bis hin zu Funktionalstilen zu beschreiben versucht hat (vgl. BEST 2005, GRZYBEK 2006). Diese frühen, zum Teil bis ins 19. Jahrhundert zurückreichenden Untersuchungen hat man Mitte des 20. Jahrhunderts dahingehend zu erweitern versucht, dass man auch die Frage nach mathematischen Verteilungsmodellen stellte, mit der sich die Wortlängenhäufigkeit im jeweils zugrunde gelegten Sprachmaterial theoretisch beschreiben lässt. Im Rahmen dieser Untersuchungen erhob man folglich zunächst die Anzahl der ein-, zwei-, drei-, usw. -silbigen Wörter und versuchte dann, sie mit diskreten Wahrscheinlichkeitsverteilungen zu modellieren.

Dabei gingen die früheren Untersuchungen aus der Mitte des 20. Jahrhunderts davon aus, dass die diskutierten Modelle sprachübergreifend gültig und damit im Prinzip von universeller Relevanz sind. Es handelte sich anfangs dabei um vergleichsweise einfache Modelle. Am einflussreichsten waren in dieser Hinsicht Untersuchungen, wie sie der russische Militärarzt Sergej G. ČEBANOV (1947) und dann vor allem der deutsche Physiker Wilhelm FUCKS (1955a, 1955b) anstellten; beide favorisierten die einfache Poisson-Verteilung, wie sie in Formel (1) dargestellt ist:

$$(1) \qquad P_x = \frac{e^{-a} a^x}{x!} \quad x = 0,1,2,\ldots$$

Es ist leicht zu sehen, dass die Poisson-Verteilung neben der konstanen Eulerschen Zahl *e* nur einen einzigen Parameter (*a*) aufweist; in Abhängigkeit von dessen Wert kann die Verteilung variieren, wobei man *a* in der Regel in Abhängigkeit vom Mittelwert oder der Varianz der jeweiligen Stichprobe interpretiert (d.h. aus der empirischen Stichprobe geschätzt) hat.

Da für die meisten Sprachen – freilich je nach linguistischer Definition (vgl. ANTIĆ / KELIH / GRZYBEK 2006) – keine 0-silbigen Wörter angenommen werden, sondern als erste Klasse in der Regel diejenige der einsilbigen Wörter

angesehen wird – wurde die Poisson-Verteilung in der Regel in ihrer 1-verschobenen Form, wie sie in (2) dargestellt ist, angewendet:

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

Die Flexibilität dieses Modells ist allerdings stark begrenzt, da es nur über einen einzigen Parameter (a) verfügt – e ist die konstante Eulersche Zahl; in der Sprach- und Textrealität hingegen haben sich Häufigkeiten von Wortlängen üblicherweise als komplexer herausgestellt. Schon allein aus diesem Grunde, vor allem aber auch bei dem Versuch, Daten von Texten mehrerer Sprachen mit einem einzigen Modell zu erfassen, hat man deshalb in der Folge zum einen Modifikationen und Erweiterungen vor allem der Poisson-Verteilung entwickelt, zum anderen aber auch weitere, hier nicht im einzelnen zu diskutierende Modelle entworfen.

Erst ab den 1980er und 1990er Jahren hat sich dann zunehmend die Einsicht durchgesetzt, dass wir es im Falle der Wortlängenhäufigkeiten (ebenso wie anderer sprachlicher Einheiten) nicht mit einem einzigen Modell zu tun haben, sondern mit einem allgemeineren Prinzip, aus dem sich dann – in Abhängigkeit von individuell zu untersuchenden Rahmenbedingungen – spezielle Modelle ableiten lassen. In der Folge haben wir es somit mit einem flexiblen System von Verteilungen zu tun. Die Grundidee besteht darin, dass die jeweils benachbarten Wahrscheinlichkeitsklassen gemäss einer einfachen Proportionalitätsbeziehung miteinander verbunden sind: $P_x \sim P_{x-1}$, d.h. die Anzahl der zweisilbigen Wörter in einem Text steht in spezifischer Relation zur Anzahl der einsilbigen Wörter dieses Textes, die Anzahl der dreisilbigen in Relation zur Anzahl der zweisilbigen, usw. Das Verhältnis zwischen den Längenklassen erweist sich dabei nicht als konstant, sondern lässt sich als Funktion verstehen: $P_x = g(x)P_{x-1}$. In Abhängigkeit davon, welche konkrete Form $g(x)$ annimmt, kommt man zu unterschiedlichen Verteilungsmodellen.

Die in der Quantitativen Sprach- und Textwissenschaft bislang diskutierten Modelle lassen sich auf einen gemeinsamen Ansatz zurückführen (WIMMER / ALTMANN 2005, 2006). Dabei können die konkreten Modelle von Sprache zu Sprache variieren, gegebenenfalls aber auch innerhalb einer Sprache in Abhängigkeit von verschiedenen Texttypen (vgl. ANTIĆ / STADLOBER / GRZYBEK / KELIH (2006). Für das Russische hat sich wiederholt herausgestellt (STITZ 1994; GIRZIG 1996; BEST / ZINENKO 2001a,b; STEINWEISS 2004), dass eine bestimmte Verallgemeinerung der Poisson-Verteilung geeignet ist, die Wortlängenhäufigkeiten in Texten verschiedener Textsorten zu modellieren.[1] Hierbei handelt es sich um die sog. Hyperpoisson-Verteilung, die in Formel (3) in ihrer 1-verschobenen Form dargestellt ist:

[1] Es ist hier nicht der Ort, diese Modelle mathematisch abzuleiten, die Komplexität dieser Modelle darzustellen oder gar mögliche Übergänge zwischen ihnen zu diskutieren; an

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

Der im Nenner stehende Ausdruck muss hier nicht im Detail erläutert werden; statt dessen mag es reichen, auf zwei wichtige Umstände hinzuweisen: zum einen auf die Tatsache, dass die Hyperpoisson-Verteilung nicht nur einen, sondern zwei Parameter (*a* und *b*) aufweist, was eine flexiblere Modellierung der Häufigkeiten erlaubt, und zum anderen, dass man für den Fall $b = 1$ die übliche Poisson-Verteilung erhält.

Im Rahmen des Grazer Wortlängen-Projekts (GRZYBEK / STADLOBER 2002, KELIH / GRZYBEK / STADLOBER 2003) konnte gezeigt werden, dass sich mit diesem Modell die Wortlängenhäufigkeiten des gesamten Spektrums von Textsorten theoretisch gut beschreiben lässt (vgl. STEINWEISS 2004): Gedichte von Aleksandr Puškin ebenso wie Prosatexte von Lev Tolstoj oder avantgardistische Texte von Daniel Charms – wobei die Existenz eines solchen einheitlichen Modells freilich nicht die Möglichkeit ausschließt, dass entweder einzelne Texte oder spezifische Textgruppen sich (sozusagen lokal) aufgrund von spezifischen Texteigenschaften gegebenenfalls mit anderen Modellen besser erfassen lassen.

Wenn damit einerseits der Nachweis erbracht ist, dass Wortlängenhäufigkeiten in russischen Texten nicht zufällig, sondern systematisch organisiert sind, und wenn damit andererseits gezeigt ist, dass sich die Häufigkeiten mit einem einheitlichen Modell erfassen lassen, so bedeutet das allerdings noch lange nicht, dass es keinerlei Variation mehr zwischen den Texten gibt. Vielmehr lassen sich aufgrund von spezifischen Analysen – in die u.a. auch die konkreten Parameterwerte als Diskriminanzfaktoren[2] eingehen – weitere Differenzierungen vornehmen.

Im vorliegenden Text soll allerdings – über die Analyse von Wortlängenhäufigkeiten hinausgehend – ein zusätzlicher und weiterführender Schritt angedacht werden. Dieser Schritt ist im Prinzip identisch mit dem oben beschriebenen Verfahren der Untersuchungen von Wortlängenhäufigkeiten, richtet sich aber nicht auf die Häufigkeiten von Wortlängen allgemein, sondern vielmehr die Häufigkeit von Wortlängenpaaren[3], also von zwei jeweils benachbarten Wortlängen. Damit ist folgendes gemeint: Haben wir zum Beispiel etwa eine Sequenz von Wortlängen wie etwa 1342534134 vorliegen, so haben wir

---

der Geschichte und an der Methodologie der Untersuchung von Wortlängenhäufigkeiten Interessierte seien deshalb auf die Fachliteratur verwiesen (vgl. GRZYBEK 2006).

2 Vgl. dazu die Arbeiten zur quantitativen Texttypologie mit der Hilfe von multivariatren Diskriminanzverfahren in GRZYBEK / KELIH (2005), GRZYBEK / KELIH / STADLOBER (2005), GRZYBEK / STADLOBER / KELIH / ANTIĆ (2005) und KELIH / ANTIĆ / GRZYBEK / STADLOBER (2005).

3 Allgemeine Studien zur Frage der Modellierung von Sequenzen finden sich in EGGHE / ROUSSEAU (1990) und EGGHE (2000).

es insgesamt mit neun Paaren zu tun (13, 34, 42, 25, 53, 34, 41, 13, 34), wobei eines dieser Paare (13) zweimal, ein anderes (34) dreimal vorkommt. Wir bekommen somit für einen gegebenen Text (oder ein Textkorpus) eine Häufigkeit von Wortlängenpaaren, und um einen ersten Versuch von deren Modellierung soll es gehen. Es handelt sich dabei um einen innovativen Ansatz, für den hier nur erste Überlegungen angestellt werden können.

Aus diesem Grunde ergibt sich für die vorliegende Untersuchung folgender Verlaufsplan: Im Anschluss an eine Darstellung des Materials, an welchem wir unsere Überlegungen präsentieren wollen, soll zunächst eine Modellierung der einfachen Wortlängenhäufigkeiten vorgenommen werden. Dies ist nicht nur „der Vollständigkeit" halber notwendig, sondern auch deshalb, weil davon auszugehen ist, dass sich zu untersuchenden Paarbildungen als Ergebnis der Überlagerung der einfachen Häufigkeiten interpretieren lassen. Die eventuell mögliche Ableitung eines derartigen mathematischen Zusammenhangs muss an anderer Stelle erfolgen; im vorliegenden Text kann in dieser Hinsicht nur ein erster Schritt getan werden, der darin besteht, dass untersucht werden kann, ob die Häufigkeiten der Wortlängenpaare einer bestimmten Regularität folgen.

Textbasis und Wortlängenhäufigkeiten

Gegenstand der vorliegenden Untersuchung sind zehn Texte von Pelevin. Titel der Texte sowie deren (in der Anzahl der Wörter berechneter) Umfang *N* sind der Tab. 1 zu entnehmen.

Tab. 1**:** Analysierte Texte und Anzahl von Wörtern (N)

No.	Text	*N*	No.	Text	*N*
1	*Buben verchnego mira*	3695	**6**	*Lunochod*	3088
2	*Den' bul'dozerista*	7515	**7**	*Proischoždenie vidov*	3619
3	*Čapaev i Pustota*	86778	**8**	*Tarzanka*	4744
4	*Chrustal'nyj mir*	6189	**9**	*Zatvornik i Šestipalyj*	9627
5	*Generation P*	55465	**10**	*Žizn' nasekomych*	42555

Veranschaulichen wir das Verfahren der Berechnung von Wortlängenhäufigkeiten und deren theoretischer Modellierung am Beispiel des längsten der zehn ausgewählten Texte, *Čapaev i pustota*. Tabelle 2 enthält in der Spalte *x[i]* die jeweilige Längenklasse und in der Spalte *f[i]* die jeweilige Vorkommenshäufigkeit der Längenklasse *x[i]*

Tab. 2: Empirische und theoretische Wortlängen-Häufigkeiten

x[i]	*f*[i]	*NP*[i]
1	28168	27938,65
2	28220	28602,95
3	17876	18012,39
4	8615	8190,63
5	2923	2914,47
6	793	851,81
7	157	211,23
8	23	45,48
9	3	10,40

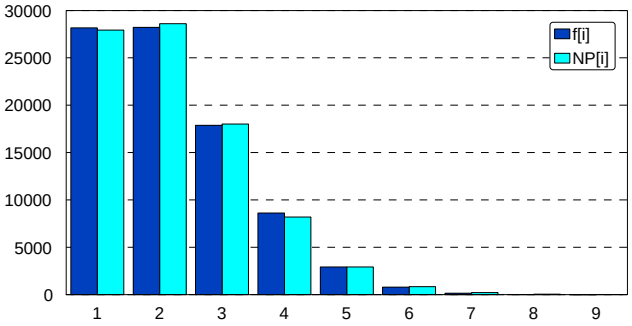

Abb 1. Anpassung an die Hyperpoisson-Verteilung

Zusätzlich enthält die Tab. 3 die sich aufgrund der Hyperpoisson-Verteilung (3) mit den Parameterwerten für $a = 1.6362$ und $b = 1.5982$ ergebenden theoretischen Werte *NP[i]*; die Ergebnisse sind in Abb. 1 anschaulich dargestellt. Die Anpassungsgüte wird mit dem Chi²-Test berechnet; in unserem Beispiel beträgt der Werte Chi² = 64.41. Dieser Wert lässt sich in Abhängigkeit von den jeweiligen Freiheitsgraden – die sich aus der Anzahl der Längenklassen minus Anzahl der geschätzten Parameter minus 1 berechnet, im gegebenen Fall also sechs beträgt – durch Bezugnahme auf die Chi²-Verteilung als Überschreitungswahrscheinlichkeit *P* interpretieren. Allerdings steigt der Chi²-Wert (und damit auch die Wahrscheinlichkeit von *P*) linear mit der Stichprobengröße an; deshalb hat es sich in sprach- und textbezogenen Untersuchungen durchgesetzt, bei Vorliegen großer Datenmengen die Bedeutsamkeit von Anpassungsergebnissen mit dem als *Chi² / N* standardisierten Diskrepanzkoeffizienten *C* zu bewerten: Bei einem Wert von $C < 0.02$ spricht man von einem guten, bei $C < 0.01$ von einem sehr guten Fit.

Im obigen Beispiel beträgt der Wert entsprechend C = 64.41 / 86778 = 0.0007 – wir haben es also mit einem ausgezeichneten Anpassungsergebnis zu

tun. Dasselbe gilt für alle zehn Stichproben, wie aus der Tab. 3 (in der die *C*-Werte für alle zehn Texte dargestellt sind) deutlich hervorgeht. Auch hier ist der Wert von $C < 0.01$ in allen Fällen unterschritten.

Tab. 3: Parameter der Hyperpoisson-Verteilung und C-Werte

Text	*a*	*b*	*C*	Text	*a*	*b*	*C*
1	1,3767	1,1779	0,0003	**6**	1,1522	1,2424	0,0044
2	1,5201	1,3061	0,0013	**7**	1,5998	1,3124	0,0012
3	1,6362	1,5982	0,0007	**8**	1,5271	1,5291	0,0007
4	1,5731	1,3418	0,0053	**9**	1,6069	1,5593	0,0043
5	1,6710	1,4878	0,0006	**10**	1,4279	1,243	0,0001

Häufigkeiten der Wortlängenpaare

Wenden wir uns auf der Basis dieser Befunde nunmehr der zweiten Fragestellung zu, die sich ja, wie beschreiben, auf die Häufigkeit von Wortlängenpaaren bezieht. Exemplifizieren wir am Beispiel des oben bereits analysierten Textes *Čapaev i pustota* das Vorgehen. Zunächst wird das Vorkommen der einzelnen Wortlängenpaare berechnet. Dabei sind verschiedene Entscheidungen zu treffen, so etwa, ob man Wortlängenpaare an Satzgrenzen bzw. über Satzgrenzen hinweg in Betracht zieht oder aber aus der Betrachtung ausschließt; im gegebenen Fall sind zunächst einmal keinerlei solcher Restriktionen eingeführt worden. Das führt zur Häufigkeit von Wortlängenpaaren, die sich am übersichtlichsten in Form einer Matrix wie in Tab. 4 anführen lassen.

Tab. 4: Häufigkeit von Wortlängenpaaren

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	Σ
1	8368	10240	5814	2658	839	208	36	5	0	28168
2	9648	8869	5790	2769	872	214	50	8	0	28220
3	6240	5318	3633	1797	648	195	40	3	1	17875
4	2814	2585	1808	910	367	111	15	4	1	8615
5	853	906	619	347	143	42	12	1	0	2923
6	191	256	164	114	43	20	2	2	1	793
7	49	41	35	18	9	3	2	0	0	157
8	3	4	12	2	2	0	0	0	0	23
9	1	1	1	0	0	0	0	0	0	3
Σ	28167	28220	17876	8615	2923	793	157	23	3	

Diese Matrix beinhaltet die den Wortlängen nach sortierten Vorkommenshäufigkeiten. Zusätzlich sind am horizontalen und vertikalen Rand die sich jeweils ergebenden Randsummen abzulesen. Mit einer solchen Matrix lassen sich eine ganze Reihe von weiterführenden Überlegungen und Berechnungen anstellen, die über den Rahmen des hier Möglichen hinausgehen und an anderer Stelle zu verfolgen sein werden. Im gegebenen Kontext wollen wir uns daher auf eine dieser Möglichkeiten beschränken, die das folgende Vorgehen beinhaltet: Ordnet man die Vorkommenshäufigkeiten der Häufigkeit nach – überführt man sie also in eine so genannte Ranghäufigkeitstabelle (vgl. Tabelle 5) – so ergeben sich für den genannten Text die folgenden Daten:

Tab. 5: Wortlängenpaarhäufigkeiten in *Čapaev i pustota*

Rang	***f(x)***	**Rang**	***f(x)***	**Rang**	***f(x)***	**Rang**	***f(x)***	**Rang**	***f(x)***	**Rang**	***f(x)***
1	10240	**12**	2658	**23**	367	**34**	50	**45**	12	**56**	2
2	9648	**13**	2585	**24**	347	**35**	49	**46**	12	**57**	2
3	8869	**14**	1808	**25**	256	**36**	43	**47**	9	**58**	2
4	8368	**15**	1797	**26**	214	**37**	42	**48**	8	**59**	2
5	6240	**16**	910	**27**	208	**38**	41	**49**	5	**60**	1
6	5814	**17**	906	**28**	195	**39**	40	**50**	4	**61**	1
7	5790	**18**	872	**29**	191	**40**	36	**51**	4	**62**	1
8	5318	**19**	853	**30**	164	**41**	35	**52**	3	**63**	1
9	3633	**20**	839	**31**	143	**42**	20	**53**	3	**64**	1
10	2814	**21**	648	**32**	114	**43**	18	**54**	3	**65**	1
11	2769	**22**	619	**33**	111	**44**	15	**55**	2	**66**	1

Die im nächsten Schritt zu verfolgende Frage richtet sich nun auf die Suche nach einem geeigneten Modell, welches die beobachtete Häufigkeit theoretisch zu beschreiben in der Lage ist. Im Prinzip impliziert das Vorgehen zwei verschiedene Wege, die einander nicht ausschließen, sondern ergänzen müssen: Denn in Ergänzung zu einem rein rechnerischen, durch iterative Prozeduren erreichbaren Modell sollte dieses Modell auch mathematisch begründbar sein und sich zum Beispiel als Überlagerung zweier Häufigkeitsverteilungen, nämlich der beiden Randverteilungen, interpretieren lassen. Beide Schritte lassen sich jedoch heuristisch voneinander trennen bzw. getrennt verfolgen. Aus diesem Grunde soll im vorliegenden Text zunächst einmal empirisch getestet werden, ob die Wortlängenpaarhäufigkeiten überhaupt einem bestimmten Modell folgen, ob es sich bei diesem Modell für die zugrunde gelegten Texte um ein einheitliches Modell handelt. Für die Untersuchung entsprechender Fragestellungen gibt es Spezialsoftware, so etwa den Altmann-Fitter, der mehr als 200 verschiedene Verteilungen in iterativen Prozeduren anpasst und auf die Eignung für das gegebene Datenmaterial testet.

Im Hinblick auf unseren Text stellt sich dabei ein überaus bekanntes Modell als geeignet heraus, nämlich die in (4) dargestellte negative Binomialverteilung (in 1-verschobener Form):

$$P_x = \binom{k+x-2}{x-1} p^k q^{x-1} \qquad x = 1, 2, 3, \ldots \tag{4}$$

Tab. 6 enthält die beobachteten sowie die sich aufgrund der negativen Binomialverteilung ergebenden theoretischen Häufigkeiten für unseren Text.

Tab. 6: Anpassungsergebnisse für die negative Binomialverteilung (*Čapaev i pustota*)

***x*[i]**	***f*[x]**	***NP*[x]**	***x*[i]**	***f*[x]**	***NP*[x]**	***x*[i]**	***f*[x]**	***NP*[x]**
1	10240	10139,41	**12**	2658	2477,49	**23**	367	435,69
2	9648	9837,49	**13**	2585	2122,26	**24**	347	371,07
3	8869	8934,95	**14**	1808	1816,27	**25**	256	315,95
4	8368	7930,65	**15**	1797	1553,16	**26**	214	268,95
5	6240	6957,32	**16**	910	1327,25	**27**	208	228,89
6	5814	6060,34	**17**	906	1133,52	**28**	195	194,76
7	5790	5253,97	**18**	872	967,55	**29**	191	165,68
8	5318	4539,39	**19**	853	825,48	**30**	164	140,92
9	3633	3911,95	**20**	839	703,98	**31**	143	119,84
10	2814	3364,5	**21**	648	600,13	**32**	114	101,9
11	2769	2889,03	**22**	619	511,42	**33**	111	86,63

Tab. 6: Fortsetzung

***x*[i]**	***f*[x]**	***NP*[x]**	***x*[i]**	***f*[x]**	***NP*[x]**	***x*[i]**	***f*[x]**	***NP*[x]**
34	50	73,64	**45**	12	12,24	**56**	2	2,02
35	49	62,59	**46**	12	10,39	**57**	2	1,71
36	43	53,19	**47**	9	8,82	**58**	2	1,45
37	42	45,2	**48**	8	7,49	**59**	2	1,23
38	41	38,4	**49**	5	6,36	**60**	1	1,04
39	40	32,62	**50**	4	5,4	**61**	1	0,89
40	36	27,71	**51**	4	4,58	**62**	1	0,75
41	35	23,54	**52**	3	3,89	**63**	1	0,64
42	20	19,99	**53**	3	3,3	**64**	1	0,54
43	18	16,98	**54**	3	2,8	**65**	1	0,46
44	15	14,42	**55**	2	2,38	**66**	1	2,56

Wie in der Abb. 2 zu sehen ist, passt die Anpassung des Modells sehr gut; dies bestätigt auch der Wert des Diskrepanzkoeffizient mit $C = 0.0084$.

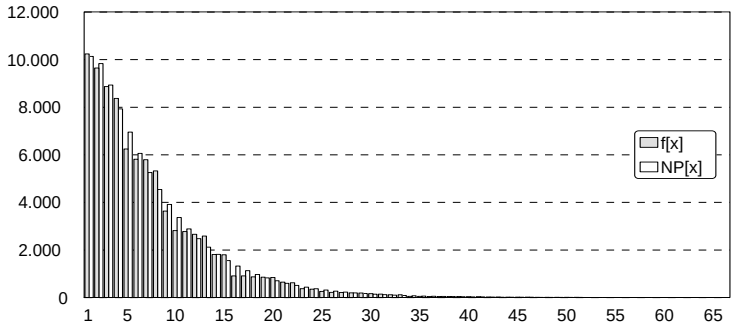

Abb. 2: Theoretische und empirische Häufigkeiten (*Čapaev i pustota*)

Die Analyse aller zehn Texte von Pelevin bestätigt weiterhin, dass die negative Binomialverteilung für all diese Texte ein geeignetes Modell ist. Tabelle 7 enthält die sich ergebenden *C*-Werte, die Auskunft über die Güte der Anpassung geben.

Tab. 7: Parameter der negativen Binomial-Verteilung und C-Werte

Text	*C*	*k*	*p*	**Text**	*C*	*k*	*p*
1	0,008	1,225	0,165	**6**	0,015	1,266	0,214
2	0,016	1,238	0,157	**7**	0,008	1,205	0,144
3	0,010	1,146	0,154	**8**	0,01	1,140	0,163
4	0,020	1,344	0,164	**9**	0,01	1,173	0,158
5	0,012	1,180	0,145	**10**	0,013	1,165	0,156

Wie zu sehen ist, liegen die Werte für alle zehn Texte bei $C < 0.02$, in drei Fällen sogar $C < 0.01$. Damit lässt sich festhalten, dass nicht nur die Häufigkeiten von Wortlängen, sondern auch die von Wortlängenpaaren einer ganz bestimmten Regularität unterliegen, die – zumindest für die von uns analysierten Texten – der negativen Binomialverteilung folgt.

Tabelle 7 enthält auch die Werte für die Parameter *k* und *p*. Hier zeigt sich, dass es weder eine Abhängigkeit eines der beiden Parameter vom Stichprobenumfang noch eine wechselseitige Abhängigkeit der beiden Parameter voneinander gibt (was jedoch an größeren Datenmengen nochmals zu überprüfen wäre). Eher scheinen die beiden Parameter relativ konstant zu sein, wie auch die Abb. 3 veranschaulicht: Versteht man beide Parameter als Zufallsvariablen und berechnet die entsprechenden 95%-Konfidenzintervalle, so zeigt sich,

dass die Schwankung der Werte relativ gering ist: $1.16 < k < 1.25$ und $0.148 < p < 0.176$.

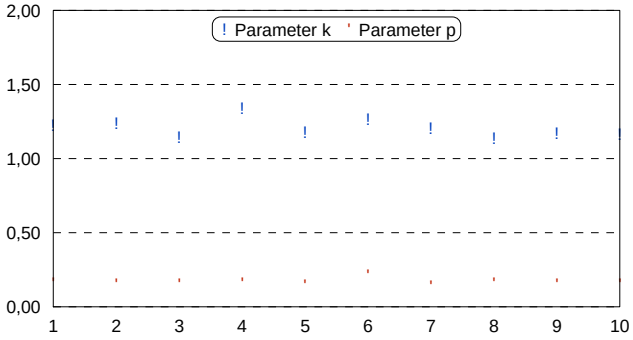

Abb. 3: Werte *k* und *p* der negativen Binomial-Verteilung

Ungeachtet der Tatsache, dass hiermit ein Modell gefunden ist, welches für die theoretische Beschreibung der Wortlängenpaare zumindest der von uns untersuchten Texte geeignet ist, scheint es sinnvoll darauf hinzuweisen, dass eine kleine Modifikation dieses Modells langfristig von Relevanz sein könnte: Diese Modifikation sieht eine Rechtsstutzung der negativen Binomialverteilung vor; das bedeutet, dass das theoretische Modell keinen unendlichen Wertebereich aufweist, sondern rechtsseitig durch die Anzahl der vorhandenen Klassen begrenzt wird. Inhaltlich ist das insofern sinnvoll, da erstens die Länge von Wörtern eine zwar im Prinzip offene, in der Realität aber begrenzte Kategorie ist, und da zudem in den einzelnen Texten in der Regel nicht alle möglichen Längenkombinationen realisiert werden. Das in (5) dargestellte rechtsgestutzte Modell weist mit dem Parameter *R* einen zusätzlichen, dem jeweiligen Inventarumfang realisierter Längenpaare entsprechenden Parameter auf:

$$(5) \qquad P_x = \binom{k+x-1}{x} q^x \qquad x = 0,1,2,\ldots,R$$

wobei

$$F(R) = \sum_{i=0}^{R} \binom{k+i-1}{i} q^i$$

In 1-verschobener Form stellt sich die rechts-gestutzte negative Binomialverteilung entsprechend wie in Formel (6) dargestellt dar:

$$(6) \qquad P_x = \frac{\binom{k+x-2}{x-1} q^{x-1}}{F(R)} \qquad x = 1,2,\ldots,R$$

wobei in diesem Fall

$$F(R) = \sum_{i=1}^{R} \binom{k+i-2}{i-1} q^{i-1}$$

Insbesondere bei kürzeren Texten könnte sich dieses Modell von besonderer Relevanz erweisen. Im Hinblick auf die zehn von uns untersuchten Texte stellt sich heraus, dass die Rechts-Stutzung der (1-verschobenen) negativen Binomialverteilung zu einer geringfügigen Verbesserung der Anpassungsergebnisse führt, wie der Tab. 8 zu entnehmen ist, in der die Werte des Diskrepanzkoeffizienten (C) mit den entsprechenden Parameterwerten für k und p aufgeführt sind.

Tab. 8: C-Werte und Parameter der (1-verschobenen) negativen Binomial-Verteilung

Text	*C*	*k*	*p*	*R*	Text	*C*	*k*	*p*	*R*
1	0,008	1,210	0,163	45	**6**	0,016	1,268	0,213	36
2	0,016	1,233	0,156	50	**7**	0,007	1,193	0,143	47
3	0,010	1,145	0,154	66	**8**	0,096	1,134	0,162	45
4	0,020	1,337	0,163	47	**9**	0,010	1,168	0,158	49
5	0,012	1,176	0,146	68	**10**	0,129	1,164	0,156	58

Ein Zusammenhang zwischen k und p ist ebenso wenig wie im Fall der nicht gestutzten Form erkennbar. Vergleicht man allerdings die Parameterwerte von k und p der 1-verschobenen negativen Binomialverteilung (4) mit denen der 1-verschobenen rechtsgestutzten negativen Binomialverteilung (6), so stellt sich heraus, dass diese hochgradig signifikant miteinander korrelieren ($r = .99$ $p < 0.001$).

Resümee und Ausblick

Im vorliegenden Text wurde die Häufigkeit von Wortlängen auf der Basis von Prosatexten des russischen Gegenwartsschriftstellers Viktor Pelevin untersucht. Es stellt sich heraus, dass die Häufigkeit der Wortlängen einer allgemeinen Regularität folgt, die auch schon an Texten anderer russischer Autoren des 19. und 20. Jahrhunderts beobachtet wurde. Diese Regularität lässt sich mit der Hyperpoisson-Verteilung theoretisch bestens modellieren. Darüber hinaus wurde im vorliegenden Text erstmals ein Versuch unternommen, über die Untersuchung von Wortlängen hinausgehend der Frage nachzugehen, ob auch die Häufigkeiten von benachbarten Wortlängen bestimmten Regularitäten unterliegen. Die erhaltenen Befunde sprechen eindeutig dafür: Die Häufigkeiten von Wortlängenpaaren folgen im Fall der untersuchten Texte der negativen Binomialverteilung. Der Frage nachzugehen, inwiefern sich dieses Modell mathematisch zur Verteilung der Wortlängenhäufigkeiten in Beziehung setzen lässt, ist eine der sich aus der vorliegenden Untersuchung ergebenden Aufgaben.

Literatur

ANTIĆ, G. / KELIH, E. / GRZYBEK, P. (2006): Zero-syllable Words in Determining Word Length. // GRZYBEK, P. (Ed.): Contributions to the science of language. Word Length Studies and Related Issues. Dordrecht, 117–157.

ANTIĆ, G. / STADLOBER, E. / GRZYBEK, P. / KELIH, E. (2006): Word Length and Frequency Distributions in Different Text Genres. // BOCK, H. H. / GAUL, W. / VICHI, M. (Eds.) (2006): Studies in Classification, Data Analysis, and Knowledge Organization. [im Druck]

BEST, K.-H. / ZINENKO, S. (2001): Wortlängenverteilungen in Briefen A.T. Twardowskis. // Göttinger Beiträge zur Sprachwissenschaft, 1, 7–19.

BEST, K.-H. / ZINENKO, S. (2001): Wortlängen in Gedichten A.T. Twardowskis. // UHLÍROVÁ, L. / WIMMER, G. / ALTMANN, G. / KÖHLER, R. (Hg.): Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Trier, 10–28.

BEST, K.-H. (2005): Wortlänge. // KÖHLER, R. / ALTMANN, G. / PIOTROWSKI, R.G. (Hg.): Quantitative Linguistik – Ein internationales Handbuch. Berlin, New York, 260-273.

ČEBANOV, S. G. (1947): On Conformity of Language Structures within the Indo-European Familiy to Poisson's Law. // Comptes Rendus (Doklady) de l'Académie des Sciences de l'URS, vol. 55, no. 2, 99–102.

EGGHE, L. (2000): The distribution of N-grams. // Scientometrics, 47, 237–252.

EGGHE, L. / ROUSSEAU, R. (1990): Introduction to Infometrics. Amsterdam.

FUCKS, W. (1955a): Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. Köln, Opladen (= Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen; 34a).

FUCKS, W. (1955b): Theorie der Wortbildung. // Mathematisch-Physikalische Semesterberichte zur Pflege des Zusammenhangs von Schule und Universität, 4, 195–212.

GIRZIG, P. (1996): Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. // BEST, K.-H. (Hg.): Glottometrika 16. The Distribution of Word and Sentence Length. Trier, 152–162.

GRZYBEK, P. (Ed.) (2006): Contributions to the Science of Language. Word Length Studies and Related Issues. Dordrecht, 15–90.

GRZYBEK, P. (2006): History and Methodology of Word Length Studies: The State of the Art. // GRZYBEK, P. (Ed.) (2006): Contributions to the Science of Language. Word Length Studies and Related Issues. Dordrecht, 15-90.

GRZYBEK, P. / STADLOBER, E. (2002): The Graz Project on Word Length (Frequencies). // Journal of Quantitative Linguistics, 9 (2), 187–192.

GRZYBEK, P. / KELIH, E. (2005): Textforschung: Empirisch! // BANKE, J. / DUMONT, B. (Hg.): Textsortenforschungen. Leipzig, 13–34.

GRZYBEK, P. / KELIH, E. / STADLOBER, E. (2005): Empirische Textsemiotik und quantitative Texttypologie. // BERNARD, J. / FIKFAK, Ju. / GRZYBEK, P. (Hg.): Text & Reality. Ljubljana, Wien, Graz, 95–120.

GRZYBEK, P. / STADLOBER, E. / KELIH, E. / ANTIĆ, G. (2005): Quantitative Text Typology: The Impact of Word Length. // WEIHS, C. / GAUL, W. (Eds.): Classification – The Ubiquitous Challenge. Heidelberg, 53–64.

KELIH, E. / GRZYBEK, P. / STADLOBER, E. (2003): Das Grazer Projekt zu Wortlängen(häufigkeiten). // Glottometrics, 6, 94–102.

KELIH, E. / ANTIĆ, G. / GRZYBEK, P. / STADLOBER, E. (2005): Classification of Author and/or Genre? The Impact of Word Length. // WEIHS, C. / GAUL, W. (Eds.): Classification – The Ubiquitous Challenge. Heidelberg, 498–505.

STEINWEISS, S. (2004): Untersuchungen zu Wortlänge und Worthäufigkeit in russischen Texten. Diplomarbeit. Graz.

STITZ, K. (1994): Untersuchungen zu den Wortlängen in deutschen und russischen Briefen des 19. Jahrhunderts. Hausarbeit im Rahmen der Ersten Staatsprüfung für das Lehramt an Gymnasien. Göttingen.

WIMMER, G. / ALTMANN, G. (2005): Unified derivation of some linguistic laws. // KÖHLER, R. / ALTMANN, G. / PIOTROVSKIJ, R.G. (Eds.): Handbook of Quantitative Linguistics, 791–807 (= Handbücher zur Sprach- und Kommunikationswissenschaft 27).

WIMMER, G. / ALTMANN, G. (2006): Towards a Unifed Derivation of Some Linguistic Laws. // GRZYBEK, P. (Ed.) (2006), 329–335.