Anzeiger für Slavische Philologie
(XXXII) 2004, 25-54

Graphemhäufigkeiten (Am Beispiel des Russischen)

Teil II: Modelle der Häufigkeitsverteilung

Peter Grzybek / Emmerich Kelih (Graz)
Gabriel Altmann (Lüdenscheid)

1. Methodologische Vorbemerkungen

Die vorliegende Studie ist der zweite Teil einer Serie von Untersuchungen zur Vorkommenshäufigkeit von Graphemen. Den Auftakt zu dieser Untersuchungsserie stellte eine historische Darstellung zur Erforschung von Graphemhäufigkeiten des Russischen dar (Grzybek/Kelih 2003), der eine Reihe allgemeiner methodologischer Bemerkungen vorausging. Als eines der Ergebnisse ging aus der synoptischen Darstellung deutlich hervor, dass es bei allen Untersuchungen nicht um die einfache Erhebung von Buchstabenhäufigkeiten an und für sich ging, sondern dass vielmehr mit allen Studien immer auch weiterführende Fragen verbunden waren, angefangen von mathematischen und methodologischen Problemen, über Fragen der Optimierung technischer Einrichtungen oder der Strukturierung von Codes und Prozessen der Informationsübertragung, bis hin zu Fragen der Textstilistik und Texttypologie. Ungeachtet der mannigfaltigen Detailfragen waren die Untersuchungen darüber hinaus aber auch immer durch ein übergeordnetes Interesse an der Vergleichbarkeit unterschiedlicher Datenerhebungen charakterisiert. Diese Frage nach der Vergleichbarkeit war dabei entweder ausgerichtet auf die „Repräsentativität" einer Stichprobe und damit primär auf die Frage der (notwendigen) Größe einer Stichprobe, oder aber auf die Bestimmung von spezifischen und generellen Charakteristiken der untersuchten Daten.

Der Frage eines einheitlichen Modells, welches verschiedenen Stichproben zugrunde liegt, soll in der vorliegenden Abhandlung im Sinne einer theoretischen Verallgemeinerung detaillierter nachgegangen werden. Dabei wollen wir nach einem einheitlichen, verschiedenen Stichproben gemeinsam zugrunde liegenden formalisierten Modell suchen. In dieser Form ist diese Frage im Hinblick auf die Häufigkeitsverteilung von Graphemen in der Ver-

gangenheit recht selten gestellt worden – und dies nicht nur im Hinblick auf russische Grapheme. Allgemein zu nennen sind in dieser Hinsicht in erster Linie Untersuchungen wie diejenigen von Sigurd (1968), Good (1969), Gusein (1988), oder Martindale et al. (1996).

Das Interesse all dieser Untersuchungen war nicht auf die Häufigkeit der jeweils einzelnen Grapheme ausgerichtet; vielmehr wurde die Frage gestellt, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, zum dritthäufigsten, usw. hat. All diese Untersuchungen haben also so genannte Rang-Häufigkeitsverteilungen untersucht – damit war das Ziel der theoretischen Modellierung die mathematische Formalisierung des Abstands zwischen den jeweiligen Häufigkeiten. Das Vorgehen hat man sich dabei wie folgt vorzustellen: Überführt man erhobene Ausgangsdaten in eine Rang-Reihenfolge, so geschieht dies üblicherweise in absteigender Reihenfolge. Wenn man sodann die jeweiligen Datenpunkte miteinander verbindet, ergibt sich charakteristischerweise kein linearer Abfall, sondern eine spezifische, monoton fallende (üblicherweise hyperbolische) Kurve. Und genau darum ist es in den genannten Untersuchungen gegangen: nämlich die genaue Form dieser Kurve zu modellieren, um so zu sehen, ob die Häufigkeiten in verschiedenen Stichproben (d.h. die spezifische Abnahme der Häufigkeiten) ein und dieselbe Form aufweisen oder nicht.

Allerdings weisen die oben genannten Untersuchungen ausnahmslos eine Reihe methodologischer Probleme auf; diese müssen hier zwar nicht im Einzelnen diskutiert werden, sollten aber zumindest in toto angesprochen werden:

1. In den Untersuchungen wird mitunter nicht konsequent zwischen der **graphematischen** und der **phonematischen** bzw. **phonologischen** Sprachebene unterschieden bzw. es wird – zumeist implizit (sei es intendiert oder aber nicht reflektiert) – davon ausgegangen, beide Spracheinheiten bzw. Repräsentationsformen würden ein und demselben Modell folgen. Letzteres ist natürlich durchaus plausibel; dennoch scheint es in einem ersten Schritt notwendig, graphematische und phonematische Ebene deutlich voneinander zu trennen. In einem zweiten Schritt lässt sich dann durchaus ein Transfer des an einer der beiden Formen als geeignet nachgewiesenen Modells auf die andere vollziehen, d.h. es lässt sich per analogiam argumentieren, was ein in der Wissenschaft gängiges Verfahren wäre. Dem müsste als nächster Schritt freilich die gesonderte Untersuchung auch der zweiten Repräsentationsform folgen, bevor gegebenenfalls schließlich die gemeinsame Betrachtung beider Formen folgen könnte.

2. Es wurde nicht konsequent unterschieden zwischen Daten, die auf Textausschnitten, auf Texten, Textkumulationen oder Textmischungen (Korpora)

beruhen – mit anderen Worten: die Bedingung der **Datenhomogenität** wurde nicht systematisch kontrolliert. Ob dies auf der Ebene der Grapheme notwendig ist, ist eine andere Frage: Es wäre durchaus nahe liegend, dass eine derart differenzierte Betrachtung bei Buchstaben nicht nötig ist – doch bedarf dies einer empirischen Überprüfung.

3. Es wurde bei der Diskussion allgemeiner Häufigkeitsmodelle bislang überwiegend mit **Kurven**, nicht mit Wahrscheinlichkeitsfunktionen gearbeitet. Auch wenn sich Kurven in letztere überführen lassen, gibt es dennoch einen wesentlichen Unterschied zwischen beiden Herangehensweisen; dieser besteht vor allen Dingen darin, dass sich bei Wahrscheinlichkeiten – im Gegensatz zu Kurvenanpassungen – die Summe der theoretisch berechneten (relativen) Häufigkeiten auf 1 belaufen muss. Zudem gilt es zu bedenken, dass die Berechnung bestimmter Eigenschaften, wie z.B. Entropie, Wiederholungsrate u.a., nur für ein Wahrscheinlichkeitssystem, nicht aber für Kurven möglich ist.

4. Die Güte des jeweils propagierten theoretischen Modells wurde auf unterschiedliche Art und Weise geprüft. Zum Teil verwendete man **Tests** für die Güte von Kurvenanpassungen (s. 3) – in diesem Fall in Form des so genannten Determinationskoeffizienten R^2. Mitunter begnügte man sich allerdings auch mit einfachen numerischen und/oder graphischen Darstellungen, in denen die beobachteten und die theoretischen Werte vergleichend gegenübergestellt wurden.

Um in Hinsicht auf die genannten Probleme ein methodologisch einheitliches Vorgehen zu erreichen, wurden für die folgenden Analysen eine Reihe von Entscheidungen getroffen:

ad 1: Es werden ausschließlich Graphemuntersuchungen durchgeführt, d.h. das zugrunde liegende Datenmaterial basiert ausschließlich auf Graphemen; entsprechend beziehen sich alle zu ziehenden Schlussfolgerungen zunächst einmal ausschließlich auf diese sprachliche Form. Eine Überprüfung, inwiefern die zu ziehenden Schlussfolgerungen auch auf die Ebene der Phoneme (oder etwa andere Sprachebenen) übertragbar sind, muss anderweitigen Analysen vorbehalten bleiben.

ad 2: Der notwendigen Datenhomogenität wird in zweierlei Hinsicht Rechnung getragen. Erstens stellen die folgenden Untersuchungen systematisch Ergebnisse zu Teiltexten, Texten, Textkumulationen und Textmischungen einander gegenüber und überprüfen diese auf die Vergleichbarkeit der Ergebnisse hin; zweitens konzentrieren sich die folgenden Analysen ausschließlich auf die Häufigkeit russischer Grapheme –

auch hier gilt, dass die Prüfung der Übertragbarkeit bzw. Erweiterbarkeit auf andere Sprachen Gegenstand weiterführender Untersuchungen sein muss.

ad 3: Es wird nicht mit Kurven, sondern mit Häufigkeitsmodellen gearbeitet. Dabei sollen alle relevanten Ansätze, die bislang zur Modellierung von Graphemhäufigkeiten vorgeschlagen worden sind, auf ihre Eignung überprüft werden. Es werden mitunter also auch solche Modelle, die Kurvenapproximationen beinhalten, zu berücksichtigen (und zum Teil) in Häufigkeitsmodelle zu überführen sein.

ad 4: Die Güte der Anpassungen soll mit statistischen Methoden überprüft werden; dazu eignet sich der sog. Chiquadrat-Anpassungstest als ein Test für die Überprüfung der Güte der Anpassung. Da der Chiquadrat-Wert allerdings linear mit der Stichprobengröße zunimmt (und man insofern bei großen Stichproben – was bei Graphemhäufigkeiten eigentlich immer der Fall ist – immer schneller mit signifikanten Abweichungen konfrontiert ist), ist es sinnvoll, den Chiquadrat-Wert mit der Stichprobengröße zu relativieren und sich auf einen Diskrepanzkoeffizienten, hier $C = \chi^2/N$, zu beziehen.

Unter diesen vereinheitlichenden Voraussetzungen wird es im folgenden möglich sein, die bislang in der Forschung diskutierten Modelle daraufhin zu testen, inwiefern sie die Häufigkeit russischer Grapheme zu modellieren geeignet sind.

2. Modelle zur Erfassung von Graphemhäufigkeiten

Allerdings scheint es angebracht, der empirischen „Eignungsprüfung“ der Modelle eine Darstellung der bislang diskutierten Modelle voranzustellen. Aus diesem Grund konzentriert sich der folgende Abschnitt auf eine Darstellung derjenigen Ranghäufigkeitsverteilungen, die bislang zur Modellierung von Graphemhäufigkeiten benutzt wurden. Da Graphemsysteme nur eine relativ begrenzte Anzahl unterschiedlicher Klassen aufweisen, ist es sinnvoll, diejenigen Verteilungen, deren Definitionsbereich nicht von $1 \ldots n$ (sondern bis unendlich) geht, auf der rechten Seite zu stutzen[1].

[1] Die Frage, inwieweit diese Modelle auch für sprachliche Einheiten höherer Ebenen geeignet sind, kann im hier gegebenen Kontext nicht weiter verfolgt werden. Ebenso werden die vielen nicht normierten Kurven, die im Zusammenhang mit der aufgewiesenen Problemstellung diskutiert worden sind, nicht näher betrachtet.

2.1. Zipf- / Zipf-Mandelbrot-Verteilung

Eines der ersten (und am häufigsten) diskutierten Modelle basiert auf den frühen Überlegungen von Zipf bzw. der Verallgemeinerung dieser Überlegungen durch Mandelbrot. Ausgehend von der Annahme, dass das Produkt aus dem Rang (r) eines Graphems und seiner Häufigkeit (f_r) eine Konstante (k) ist, nimmt die entsprechende Gleichung die Form $f_r \times r = k$ an, die im Hinblick auf die theoretische Berechnung der Häufigkeit als

$$(1) \quad f_r = \frac{k}{r}, \; r = 1,2,3,...$$

dargestellt wird. Da diese Formel jedoch keine Verteilung darstellt – weil, wie oben gesagt wurde, die Summe der relativen Häufigkeiten nicht auf 1 hinausläuft (die harmonische Reihe konvergiert nicht, k ist keine Normierungskonstante) – wurde sie um einen Parameter bereichert. Das daraus resultierende Verteilungsmodell wird üblicherweise als Zipf-Verteilung oder auch als Zeta-Verteilung[2] (Wimmer/Altmann 1999: 664f.) bezeichnet:

$$(2) \quad P_r = \frac{k}{r^a}, \quad r = 1,2,3,..., \quad a > 1, \;\; k^{-1} = \sum_{j=1}^{\infty} \frac{1}{j^a}$$

Die Erweiterung bzw. Verallgemeinerung der ursprünglichen Überlegungen von Zipf durch Mandelbrot beinhaltete eine flexiblere Formel mit einem weiteren Parameter: Ausgehend von der Ausgangsgleichung $f_r \times r = k$ stellt diese sich dar als $f_r \times (b + r)^a = k$, woraus sich im Hinblick auf die Berechnung der Häufigkeit leicht ergibt

$$(3) \quad f_r = \frac{k}{(b+r)^a}, \quad r = 1,2,3,...$$

Das daraus resultierende Verteilungsmodell wird üblicherweise als Zipf-Mandelbrot-Verteilung bezeichnet (Wimmer/Altmann 1999: 666):

$$(4) \quad P_r = \frac{k}{(b+r)^a}, \quad r = 1,2,3,..., \; a > 1, \;\; b > -1, \;\; k^{-1} = \sum_{j=1}^{\infty} \frac{1}{(b+j)^a}$$

---

[2] Den Namen Zeta-Verteilung hat sie deshalb erhalten, weil in der Gleichung (2) $k^{-1} = \zeta(a)$ die Riemannsche Zeta-Funktion ist; sie hat aber viele weitere Namen wie diskrete Pareto-Verteilung, Joos-Modell, Riemannsche Zeta-Verteilung, Zipf-Estoup-Verteilung, Zipf'sches Gesetz u.a. (vgl. Wimmer/Altmann 1999).

Wie der Formel (4) zu entnehmen ist, handelt es sich um eine Verteilung mit unendlichem Definitionsbereich; stutzt man sie auf der rechten Seite, bekommt man aus (4) die Verteilung

$$(5) \quad P_r = \frac{k}{(b+r)^a}, \quad r = 1,2,3,...,n, \quad a \in \Re, \; b > -1, \; k^{-1} = \sum_{j=1}^{n} \frac{1}{(b+j)^a}$$

2.2. Geometrische Verteilung

Ein weiteres Modell, das wiederholt im Zusammenhang mit Graphemhäufigkeiten diskutiert worden ist, basiert auf der so genannten geometrischen Folge, die aus den Gliedern

$$aq^0, aq^1, aq^2, aq^3, ..., aq^{n-1}, \ldots$$

besteht. In Analogie zu den Überlegungen von Zipf (s.o.) lässt sich hieraus die Funktion

$$(6) \quad f_r = a \cdot q^r, \quad r = 0,1,2,...$$

aufstellen, die im Zusammenhang mit Graphemhäufigkeiten u.a. auch bei Sigurd (1968) oder Martindale et al. (1996) diskutiert wird. Da bei Rangierungen der erste Rang üblicherweise als „1“ (und nicht als „0“) bezeichnet wird und die Funktion somit mit $r = 1$ anfängt, hat man in der Regel entweder die 1-verschobene Form

$$(7) \quad f_r = a \cdot q^{r-1} \qquad r = 1,2,3,...,$$

verwendet, welche die Verteilung

$$(8) \quad P_r = pq^{r-1} \qquad r = 1,2,3,..., \qquad 0 < q < 1, p = 1-q$$

ergibt, oder die 1-verschobene rechts-gestutzte Form:

$$(9) \quad P_r = \frac{(1-q)q^{r-1}}{1-q^n}, \qquad r = 1,2,...,n, \; 0 < q < 1$$

2.3. Good-Verteilung

Drei der obigen Verteilungen – nämlich (2), (8) und (9) – lassen sich als Spezialfälle einer weiteren Verteilung verstehen, die ebenfalls im Zusammenhang mit Graphemhäufigkeiten diskutiert worden ist. Dabei handelt es sich um die sog. Good-Verteilung. Da Good mehrere Verteilungsmodelle entwickelt hat, wird sie in der Fachliteratur als Good-1-Verteilung bezeichnet (Wimmer / Altmann 1999: 219f.). Letztere wird von Martindale et al. (1996) ins Spiel gebracht, und zwar in Form der folgenden Gleichung:

$$(10) \qquad P_r = \frac{a}{r^b} \cdot c^r, \quad r = 1, 2, ..., n$$

Hierbei stellt a eine Normierungskonstante dar, die bewirkt, dass die Summe der relativen Wahrscheinlichkeiten 1 ist:

$$a^{-1} = \sum_{j=1}^{n} \frac{c^j}{j^b} \quad a^{-1} = \sum_{j=1}^{n} \frac{c^j}{j^b} \cdot$$

Der Zusammenhang zwischen der in (10) dargestellten Good-1-Verteilung zu den oben genannten Verteilungen ergibt sich wie folgt. Setzt man in (10)

a) $b = 0$, $c = q$, $a = p/q$, $n \rightarrow \infty$, so erhält man Verteilung (8);
b) für $a = (1\text{-}q) /[q(1\text{-}q^n)]$, $b = 0$, $c = q$, erhält man die rechts-gestutzte geometrische Verteilung (9);
c) mit $a = k$, $b = a$, $c = 1$ und $n \rightarrow \infty$ erhält man die Zeta-Verteilung (2).

Um alle obigen Verteilungen einheitlich darstellen zu können (vgl. Zörnig, Altmann 1995), führen wir hier die Lerchsche zeta-Funktion ein. Sei

$$(6) \qquad \Phi(p, b, a) = \sum_{j=1}^{\infty} \frac{p^j}{(b + j)^a} \cdot$$

Mehrere Summen lassen sich mit dieser (in Wimmer-Altmann 1999 etwas anders definierten) Funktion symbolisch darstellen. Die Summation bis n kann man folgendermaßen ausdrücken:

$$\sum_{j=1}^{n} \frac{p^j}{(b+j)^a} = \sum_{j=1}^{\infty} \frac{p^j}{(b+j)^a} - \sum_{j=n+1}^{\infty} \frac{p^j}{(b+j)^a}$$

$$= \Phi(p,b,a) - \sum_{j=1}^{\infty} \frac{p^{j+n}}{(b+j+n)^a}$$

$$= \Phi(p,b,a) - p^n \sum_{j=1}^{\infty} \frac{p^j}{(b+j+n)^a}$$

$$= \Phi(p,b,a) - p^n \Phi(p,b+n,a)$$

Stutzt man die Zeta-Verteilung (2) rechts, so erhält man als Normierungskonstante der rechts-gestutzten Zeta-Verteilung entsprechend:

(2a) $$P_r = \frac{1}{r^a[\Phi(1,0,a) - \Phi(1,n,a)]}, \quad r = 1,2,...,n$$

die man als rechts gestutzte Zeta-Verteilung bezeichnet (Wimmer / Altmann 1999: 577f.). Formel (5), die rechts gestutzte Zipf-Mandelbrot-Verteilung (vgl. Mandelbrot 1953), kann man dann als

(5a) $$P_r = \frac{1}{(b+r)^a[\Phi(1,b,a) - \Phi(1,b+n,a)]}, \quad r = 1,2,...,n$$

schreiben. Weiterhin ergibt sich die rechts gestutzte geometrische Verteilung (9) als

(9a) $$P_r = \frac{q^{r-1}}{[\Phi(q,0,0) - \Phi(q,n,0)]}, \quad r = 1,2,...,n$$

Die Good-Verteilung (10) schließlich kann man darstellen als

(10a) $$P_r = \frac{c^r}{r^b[\Phi(c,0,b) - \Phi(c,n,b)]}, \quad r = 1,2,...,n.$$

Damit stellt sich heraus, dass die oben erwähnten Verteilungsmodelle sich auf ein gemeinsames, übergeordnetes Modell zurückführen lassen, als dessen Spezialfälle sich die aufgeführten Verteilungen erweisen. In dieser Richtung ist somit die Lerch-Verteilung, die sich mit Hilfe von (6) ergibt (vgl. Zörnig/ Altmann 1995), das allgemeinste Modell, das allerdings bislang noch nicht im Hinblick auf Graphemhäufigkeiten getestet wurde.

2.4. Synergetische Verallgemeinerungen

Ein weiteres Verteilungsmodell ist von Martindale et al. (1996) ins Spiel gebracht worden, wenn auch nur in Form einer Kurvenapproximation diskutiert worden (s.u.). Abgesehen davon, dass es also in eine Wahrscheinlichkeitsfunktion überführt werden muss, ist es geboten, dieses Modell in einem etwas weiteren Kontext zu verankern, zumal

(a) das Modell auch in umfangreicher Fachliteratur wie dem *Thesaurus univariater diskreter Verteilungen* von Wimmer/Altmann (1999) keine Berücksichtigung findet und auch in entsprechender Spezialsoftware (Altmann-Fitter) nicht enthalten ist, und
(b) die hinter diesem Modell steckenden Überlegungen bislang – zumindest im Hinblick auf die Untersuchung von Graphemhäufigkeiten – noch nie systematisiert worden sind.

Zum Zwecke der notwendigen Systematisierung ist es sinnvoll, auf Grundannahmen der synergetischen Linguistik zu rekurrieren. In diesem Zusammenhang wird bei Verteilungsproblemen im allgemeinen angenommen, dass die Wahrscheinlichkeit einer bestimmten Klasse mit der Ausprägung x oder dem Rang r sich proportional zu der jeweils niedrigeren Klasse (also x–1 bzw. r–1) entwickelt (vgl. Altmann / Köhler 1996). Von diesem allgemeinen Ansatz ausgehend kann man also die Differenzengleichung

$$(11) \qquad P_x = g(x)P_{x-1}$$

aufstellen, deren konkrete Lösung von der jeweiligen Funktion $g(x)$ abhängt. Bereits einfache Funktionen haben in den vergangenen Jahren bei der Untersuchung der Vorkommenshäufigkeit verschiedener sprachlicher Einheiten zu überzeugenden Resultaten geführt. Dabei war $g(x)$ üblicherweise eine einfache gebrochene rationale Funktion, bei deren qualitativer Interpretation man davon ausging, dass sie im Zähler die „Kräfte" des Sprechers, im Nenner die regulierenden „Kräfte" des Hörers enthielt. Wimmer / Altmann (2003b) verallgemeinerten diesen Ansatz nunmehr in zwei Richtungen:

(a) Im ersten Fall gibt es nur eine einzige unabhängige Variable, nämlich x; alle anderen werden als ceteris paribus betrachtet; da sie aber nicht alle gleich sind und keine Konstante darstellen, werden sie gewichtet, und zwar als b/x^2, c/x^3, ...
(b) Im zweiten Fall gibt es mehrere unabhängige Variablen, die alle explizit berücksichtigt werden, wodurch sich partielle Differenzen- und Differentialgleichungen ergeben.

Im folgenden wird nur Fall (a) in Betracht gezogen (wobei man in der Linguistik hier zunächst nur bis zu drei, höchstens vier Gliedern gekommen ist). Einige davon sind:

$$(12) \quad P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2}\right) P_{x-1}$$

$$(13) \quad P_x = \left(1 + a_0 + \frac{a_1}{x + b_1} + \frac{a_2}{(x + b_2)^2}\right) P_{x-1}$$

$$(14) \quad P_x = \left(1 + a_0 + \frac{a_1}{(x + b_1)^{c_1}} + \frac{a_2}{(x + b_2)^{c_2}}\right) P_{x-1}$$

usw., deren Lösungen vergleichsweise einfach sind. Man sieht, dass (13) ein Spezialfall von (14), und (12) ein Spezialfall von (13) ist. Für die geometrische Verteilung (s.o.) lässt sich zeigen, dass sie ein Spezialfall von (12) ist, nämlich dann, wenn $a_0 = -p$, $a_1 = a_2 = 0$, $1 - p = q$. In diesem Fall bekommt man die übliche, die verschobene oder die rechts-gestutzte geometrische Verteilung, je nach Art der Summierung.

Um auch die anderen oben genannten Verteilungen (1)–(10) unter dieses allgemeine Modell zu bringen, muss man allerdings eine weitere Verallgemeinerung des Wimmer-Altmannschen Ansatzes durchführen, nämlich

$$(15) \quad P_x = \left(1 + a_0 + \frac{a_1}{x + b}\right)^c P_{x-1}$$

wobei die ersten drei Glieder der allgemeinen Formel (nach Wimmer /Altmann 2003) zur gemeinsamen Erfassung der genannten Verteilungen ausreichen. Setzt man in (15) $a_0 = 0$, $a_1 = -1$, dann erhält man die Rekursionsformel

$$(16) \quad P_x = \left(1 - \frac{1}{x + b}\right)^c P_{x-1},$$

deren schrittweise Lösung die Zipf-Mandelbrot-Verteilung liefert. Man beachte

$$P_2 = \left(1 - \frac{1}{2 + b}\right)^c P_1 = \left(\frac{b + 1}{b + 2}\right)^c P_1$$

$$P_3 = \left(1 - \frac{1}{3 + b}\right)^c P_2 = \left(\frac{b + 2}{b + 3}\right)^c P_2 = \left(\frac{b + 2}{b + 3}\right)^c \left(\frac{b + 1}{b + 2}\right)^c P_1$$

……

$$P_x = \left(1 - \frac{1}{x+b}\right)^c P_{x-1} = \left(\frac{b+x-1}{b+x}\right)^c \left(\frac{b+x-2}{b+x-1}\right) P_{x-2} = \dots$$
$$\dots = \left(\frac{b+x-1}{b+x}\right)^c \dots \left(\frac{b+1}{b+2}\right)^c P_1 = \left(\frac{b+1}{x+b}\right)^c P_1$$

Setzt man hier $(b+1)^c P_1 = k$, dann erhält man die in Formel (4) bzw. (5) dargestellte Zipf-Mandelbrot-Verteilung; auf ähnliche Weise bekommt man die oben aufgeführten Spezialfälle der Zipf-Mandelbrot-Verteilung.

Die Differenzengleichung (11) – die sozusagen den „Blick von oben nach unten" zeigt – ist eher für Entitäten mit großem Inventarumfang geeignet. Ist der Inventarumfang hingegen klein, dann müssen sich alle Häufigkeiten untereinander so ausbalancieren, dass sie zueinander in einem solchen Verhältnis stehen, welches „vorgeschriebene" Charakteristika (wie z.B. Entropie oder Wiederholungsrate) liefert. Das aber bedeutet, dass die Abhängigkeit nicht nur „nach unten", sondern auch „nach oben" besteht. Dies lässt sich besonders gut mit Hilfe von Partialsummenverteilungen bewerkstelligen, die man folgendermaßen definieren kann: Sei $\{P^*_j\}$ eine gegebene Wahrscheinlichkeitsfunktion (die man als ‚Ausgangsverteilung' bezeichnen kann); dann ergibt sich daraus eine neue Verteilung als

$$P_x = C \sum_{j \geq x+k} f(P_j^*). \tag{17}$$

Wimmer/Altmann (2000) haben allgemein verschiedene Schemata mathematisch analysiert und im Anschluss daran für linguistische Zwecke dargestellt (Wimmer / Altmann 2001). Von den dort aufgezeigten Optionen wurde im Zusammenhang mit Graphemanalysen – freilich ohne Bezug auf dieses Schema – bisher nur ein einziger Fall verwendet, der sich nach Wimmer / Altmann (2001: 285) dem Schema II zuordnen lässt, nämlich die Partialsumme der diskreten Rechteckverteilungen (vgl. Good 1969, Gusein-Zade 1988, Martindale et al. 1996). Das entsprechende kombinatorische Schema wurde bereits von Whitworth (1901: 207f.) benutzt; es wird in der Literatur auch als „broken stick distribution", als „distribution of ordered random intervals", oder als „MacArthur distribution" bezeichnet.[3] Diese Verteilung, die hier als **Whitworth-Verteilung** bezeichnet werden soll, entsteht folgendermaßen:

Sei $P_j^* = \frac{1}{n}, \quad j = 1,2,...,n$, d.h. die diskrete Rechteckverteilung, und wir setzen sie in das Schema

[3] Unter dem letzten Namen wird sie auch in Wimmer / Altmann (1999: 401) in etwas anderer (umgekehrter) Form dargestellt; dort findet sich auch weitere Literatur.

$$(18) \quad P_x = C\sum_{j\geq x}\frac{P_j^*}{j}, \quad x = 1,2,3,...$$

ein. So erhält man sehr einfach

$$(19) \quad P_x = \frac{1}{n}\sum_{j=x}^{n}\frac{1}{j}, \quad x = 1,2,3,...,n$$

wo $C = 1$ ist. Wenn man die einzelnen P_x explizit ausschreibt und addiert, so sieht man, dass die Summe 1 ergibt. Diese Verteilung hat den Vorteil, dass sie nur einen einzigen Parameter (n) hat, der sich aus dem Inventarumfang ergibt (und insofern leicht zu interpretieren ist).

2.5. Negative hypergeometrische Verteilung

Ein weiteres Verteilungsmodell, das es im hier gegebenen Zusammenhang zu diskutieren gilt, ist die sog. negative hypergeometrische Verteilung (Wimmer/ Altmann 1999: 465ff.), die u.a. auch unter der Bezeichnung Beta-Binomial-Verteilung bekannt ist. Dieses Modell ist wiederholt allgemein zur Modellierung verschiedener Ranghäufigkeiten eingesetzt worden (vgl. z.B. Köhler/ Martináková 1998, Wimmer/Altmann 2001, Wimmer/Wimmerová 2003); insbesondere auch zur Modellierung der Graphemhäufigkeit in einem ausgewählten Puškin-Text ist es verwendet worden (Grzybek 2001), ohne dass allerdings in dieser Richtung systematische Studien durchgeführt worden wären.

Man kann diese Verteilung auf verschiedene Weisen ableiten; für unsere Zwecke reicht es, sie aus dem Ansatz (14) herzuleiten. Setzt man hier nämlich

$$a_0 = b_2 = 0,$$
$$a_1 = (-K + M + 1)(K + n - 1)/(-K + M - n),$$
$$a_2 = (n+1)(M-1)/(K - M + n),$$
$$b_1 = -K + M - n$$
$$b_2 = 0, K > M \geq 0, n \in \{0,1,\ldots\}, c_1 = c_2 = 1$$

so erhält man

$$(20) \quad P_x = \frac{(M + x - 1)(n - x + 1)}{x\left(K - M + n - x\right)}P_{x-1},$$

woraus die negative hypergeometrische Verteilung resultiert:

$$(21) \quad P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}} \qquad \begin{array}{l} x = 0,1,2,\ldots,n, \\ K > M > 0;\ n \in \{1,2,\ldots\} \end{array}$$

Für Rangierungszwecke muss man diese Verteilung um einen Schritt nach rechts verschieben und erhält so die 1-verschobene negative hypergeometrische Verteilung.

$$(22) \quad P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}} \qquad \begin{array}{l} x = 1,2,\ldots,\ n+1 \\ K > M > 0;\ n \in \{1,2,\ldots\} \end{array}$$

Stutzt man sie im Punkt 0, erhält man die positive negative hypergeometrische Verteilung:

$$(23) \quad P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n} - \binom{K-M+n-1}{n}}, \quad x = 1,2,\ldots,n$$

Auch zwischen dieser Verteilung und der Whitworth-Verteilung lässt sich ein Zusammenhang herstellen: Setzt man nämlich in (23) $K = 2$ und $M = 1$ ein, so erhält man die diskrete Rechteckverteilung; bildet man sodann Partialsummen wie oben beschrieben, stellt sich die Whitworth-Verteilung als ein Spezialfall der partial summierten negativen hypergeometrischen Verteilung dar.

3. Empirische Überprüfung der Modelle

3.1. Text- und Datenbasis

Bei der empirischen Überprüfung der oben dargestellten Modelle an russischen Graphemen soll, wie oben bereits angesprochen wurde, vor allem die Frage der Datenhomogenität systematisch kontrolliert werden. Es sind zwar auf der Ebene der Grapheme nicht unbedingt durch die Verletzung der Datenhomogenität bedingte Inkonsistenzen zu erwarten, doch soll eine entsprechend systematische Kontrolle dieses Faktors auf jeden Fall gewährleistet sein.

Aus diesem Grunde sollen die oben referierten Modelle an verschiedenem (jedoch ausschließlich russischem) Datenmaterial getestet werden; im Detail handelt es sich um die folgenden Dateneigenschaften:

a. In einer ersten Textgruppe handelt es sich um ***vollständige Texte***; um dabei keiner spezifischen Definition von „Text" folgen zu müssen, werden unter vollständigen Texten sowohl in sich abgeschlossene Kapitel eines Romans als auch vollständige Romane herangezogen. Überwiegend handelt es sich um literarische (und zwar ebenso prosaische wie poetische und dramatische) Texte; dennoch sind zum Zwecke des Vergleichs auch technische Texte berücksichtigt.
b. In einer zweiten Textgruppe handelt es sich nicht um vollständige Texte, sondern um ***Textausschnitte***, ***Textkumulationen***, ***Textmischungen***, und ein vollständiges ***Textkorpus***. Textausschnitte stellen willkürlich ausgewählte Textpassagen dar, so z.B. einzelne Kapitel eines Textes, bestimmte Verse oder Zeilen eines Kapitels o.ä.; unter Text-Kumulationen sind sukzessiv addierte Kapitel eines vollständigen Textes zu verstehen, die im letzten Schritt dem Gesamttext entsprechen; Textmischungen beinhalten die Kombination willkürlich ausgewählter Texte (oder auch Teile verschiedener Texte); die Mischung aller hier untersuchten Texte geht schließlich in ein Textkorpus ein, das im gegebenen Fall immerhin einen Umfang von mehr als 8.5 Millionen Buchstabenvorkommnissen aufweist.

Tab. 1 stellt eine Übersicht über die vollständigen Texte und Textkumulationen dar. Im Anschluss an die Nummer des Textes – auf die auch in den einzelnen Analysen Bezug genommen werden wird – findet sich eine Angabe zum Autor bzw. zur Quelle des Textes, sodann die Bezeichnung des Textes, der Textstatus, eine Kurzbezeichnung des Textes und schließlich der Umfang des Textes (in der Anzahl der Buchstabenvorkommnisse).

Tab. 1: Text- und Datenbasis: Vollständige Gesamttexte und Textkumulierungen

Nr.	Autor	Text	Kapitel	Kürzel	N
1	A.S. Puškin	*Evgenij Onegin*	1	ASP-EO 1	15830
2			2	ASP-EO 2	11544
3			3	ASP-EO 3	13597
4			4	ASP-EO 4	12475
5			5	ASP-EO 5	12018
6			6	ASP-EO 6	12742
7			7	ASP-EO 7	15180
8			8	ASP-EO 8	15864
9			1-2	ASP-EO 1-2	27374
10			1-3	ASP-EO 1-3	40971
11			1-4	ASP-EO 1-4	53446
12			1-5	ASP-EO 1-5	65464
13			1-6	ASP-EO 1-6	78206
14			1-7	ASP-EO 1-7	93386
15			Gesamttext	ASP-EO 1-8	109250
16	L.N. Tolstoj	*Anna Karenina*	Gesamttext	LNT-AK	1336483
17		*Otročestvo*	Gesamttext	LNT-O	113954
18	F.M. Dostojevskij	*Prestuplenie i nakazanie*	Gesamttext	FMD-PN	837885
19		*Zapiski iz podpol'ja*	Gesamttext	FMD-ZAP	188249
20	A.P. Čechov	*Čajka*	Gesamttext*	APČ-Č	145735
21		*Djadja Vanja*	Gesamttext*	APČ-DV	60871
22	M. Gor'kij	*Mat'*	Gesamttext	MG-MA	433177
23		*Na dne*	Gesamttext*	MG-ND	76039
24	http://www.rusmet.ru/	*Ural'skij rynok metallov*	technischer Text	UR	8061
25	http://www.phyton.ru/	*Instrumental'nye sredstva […]*	technischer Text	IN	18711
		* Die hier angeführten, mit einem Asterisk gekennzeichneten Dramentexte beinhalten sämtliche Regieanweisungen, Personennennungen, etc.			

Auf dieselbe Art und Weise ist auch die Tab. 2 aufgebaut, der die entsprechenden Angaben für die Textmischungen, die Textausschnitte und das Gesamtkorpus[4] zu entnehmen sind:

Tab. 2: Text- und Datenbasis: Textmischungen, Textausschnitte und Gesamtkorpus

Nr.	Autor	Text	Kapitel	Kürzel	N
26	A.S. Puškin	*Evgenij Onegin*	Kap. 1 & 8	ASP-EO1+8	31694
27	L.N. Tolstoj	*Anna Karenina*	Teil 8 (Kap. 18) & Teil 1 (Kap. 1)	LNT-AK8+1	7720
28	F.M. Dostojevskij	*Prestuplenie i nakazanie*	Teil 1 (Kap. 1) & Teil 6 (Kap. 8)	FMD-PN1+6	29498
29	A.S. Puškin & L.N. Tolstoj	*Evgenij Onegin & Anna Karenina*	Gesamttexte	ASP+LNT	1445733
30	A.S. Puškin & F.M. Dostojevskij	*Evgenij Onegin & Prestuplenie i nakazanie*	Gesamttexte	ASP+FMD	947135
31	A.S. Puškin & Text 24	*Evgenij Onegin & Text 24*	Gesamttexte	ASP+UR	117311
32	L.N. Tolstoj & Text 24	*Anna Karenina & Text 24*	Gesamttexte	LNT+UR	1344544
33	F.M. Dostojevskij & Text 25	*Prestuplenie i nakazanie & Text 25*	Gesamttexte	FMD+IN	856596
34	M. Gor'kij und Text 25	*Na dne & Text 25*	Gesamttexte	MG+IN	95312
35	Puškin, A.S.	*Evgenij Onegin*	Kap. 5, jeweils Verse 1-5	ASP1-5	4323
36	F.M. Dostojevskij	*Prestuplenie i nakazanie*	Epilog, jede zweite Zeile	FMD-2	14464
37	L.N. Tolstoj	*Anna Karenina*	Teil 4 (Kap. 1-5), jede 4 Zeile	LNT-4	7141
38	Gesamtkorpus			GK	8697983

3.2. Ergebnisse

Schauen wir uns im folgenden die Ergebnisse für oben diskutierten Verteilungsmodelle im einzelnen an. Die Grundidee besteht darin, die Parameter der verschiedenen Gleichungen (d.h. die Variablen und Konstanten), für jeden einzelnen Datensatz so zu berechnen, dass die Abweichungen zwischen den empirischen und den theoretischen Werten minimal werden. Für jedes

[4] Die Bezeichnung ‚Gesamtkorpus' ist gewissermaßen irreführend, weil eine Reihe von Texten bzw. Textausschnitten, -kumulationen, und -mischungen zwei- oder mehrfach Eingang in den als ‚Gesamtkorpus' bezeichneten Datensatz #38 gefunden haben.

einzelne Verteilungsmodell können die Parameter also variieren, ohne dass sich die allgemeine Formel dabei ändert. Die Güte einer solchen Anpassung, auf deren Basis sich die theoretischen (geschätzten) Werte ergeben, wird in der weiteren Folge dann in der Regel mit einem so genannten χ^2-Anpassungstest geprüft. Da dieser Test jedoch bei großen Stichproben (mit denen man bei sprachlichem Material, zumal bei Graphemhäufigkeiten, in der Regel zu tun hat), relativ schnell signifikant wird, verwendet man bei Stichproben mit großem N statt dessen auch den als χ^2 / N berechneten Diskrepanzkoeffizienten C; dieser wird bei $C < 0.02$ als Indiz einer guten, bei $C < 0.01$ als Indiz einer sehr guten Anpassung angesehen – in diesem Fall ist somit davon auszugehen, dass die theoretische Berechnung geeignet ist, die empirisch ermittelten Werte in dem gegebenen Modell zu erfassen.

Insgesamt ist man dann natürlich bestrebt, einem solchen Modell den Vorzug zu geben, das nicht nur auf einen guten Anpassungswert kommt, sondern auch möglichst wenig Parameter aufweist, da ein solches Modell in der Regel leichter interpretierbar ist, so dass der Weg von der quantitativen zur qualitativen Analyse leichter beschritten werden kann.

Die weiter unten folgenden Tabellen mit den Ergebnissen der Anpassungen enthalten neben der Textnummer und dem jeweiligen Kürzel des Textes (s.o.) den sich aus der Anpassung der Verteilungsmodelle ergebenden Wert für den bzw. die Parameter der jeweiligen Verteilung, den χ^2-Wert mit der dazugehörigen Anzahl der Freiheitsgrade (*FG*), sowie den Wert des Diskrepanzkoeffizienten *C*.

Veranschaulichen wir das Vorgehen zunächst an einem ausgewählten Beispiel und passen dazu die rechts-gestutzte Zeta-Verteilung an die Daten des Gesamtkorpus an. Tab. 3 enthält neben den Rängen 1 bis 32 die absoluten Häufigkeiten *f(i)* in absteigender Reihenfolge. Der Parameter $R = 32$ berechnet sich unmittelbar aus dem Inventarumfang; für den Parameter a, der sich auf verschiedene Arten und Weisen schätzen und dann in iterativen Verfahren optimieren lässt, ergibt sich im gegebenen Fall $a = 0{,}712116987825403$. Setzt man diese Werte in die Formel (2b) ein, ergeben sich die theoretischen Werte *NP(i)* – vgl. auch Formel (2a):

$$(2) \qquad P_r = \frac{x^{-a}}{F(R)}, \quad r = 1,2,3,...,R, \quad a \in \mathbb{R}, \quad R \in \mathbb{N}, \quad F(R) = \sum_{i=1}^{R} i^{-a}$$

Tab. 3:

i	*f*(i)	*NP*(i)	i	*f*(i)	*NP*(i)
1	982048	1328881,02	**17**	181684	176710,57
2	763584	811180,54	**18**	163449	169662,25
3	701891	607742,97	**19**	156929	163254,04
4	593949	495163,87	**20**	151944	157398,48
5	563851	422413,40	**21**	146832	152023,69
6	532783	370980,75	**22**	138459	147070,00
7	456610	332412,56	**23**	101994	142487,44
8	423657	302259,79	**24**	93156	138233,79
9	403285	277941,76	**25**	77999	134273,18
10	353818	257851,17	**26**	69870	130574,85
11	295548	240931,00	**27**	54464	127112,32
12	272216	226455,46	**28**	30854	123862,62
13	262459	213908,53	**29**	24421	120805,76
14	248196	202912,52	**30**	22314	117924,21
15	222221	193184,17	**31**	9578	115202,55
16	195629	184506,56	**32**	2257	112627,18
	a =0,7121			χ^2 =1031004,65	
	R =32			FG =29	
				C =0,1185	

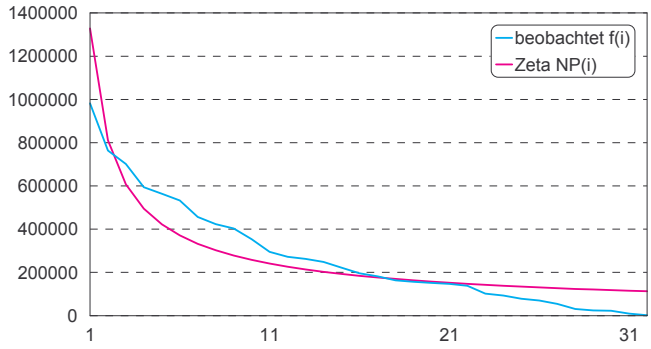


Abb. 1

Anpassung der rechts-gestutzten Zeta-Verteilung
(Gesamtkorpus)

Wie den Werten in Tab. 3 und ihrer Veranschaulichung in Abb. 1 zu entnehmen ist, stellt die Zeta-Verteilung für die Daten des Gesamtkorpus kein gutes Modell dar, was sich auch im Wert des Diskrepanzkoeffizienten $C = 0.1185$ klar ausdrückt. Diese Einschätzung gilt allerdings nicht nur für das gesamte Korpus, sondern für alle anderen Einzelstichproben in gleicher Weise. Tab. 4a/b enthält für alle einzelnen Datensätze sowohl die Ergebnisse für die rechts-gestutzte Zeta-Verteilung (2a/b) als auch für die Zipf Mandelbrot-Verteilung (5).

Deutlich ist zu sehen, dass beide Verteilungen, die in der Vergangenheit wiederholt in Betracht gezogen worden sind, sich nicht für die Modellierung der Ranghäufigkeit russischer Grapheme eignen: Bei der Zeta-Verteilung liegen die Werte des Diskrepanzkoeffizienten für die einzelnen Datensätze im Intervall von $0.1664 \geq C \geq 0.0995$, für das gesamte Korpus beträgt $C = 0.1185$, und insgesamt kommt nicht eine einzige Stichprobe auf einen Wert von $C < 0.02$. Und auch bei der Zipf-Mandelbrot-Verteilung, die mit drei Parametern (a,b,n) einen Parameter mehr als die Zeta-Verteilung hat, belegen die Ergebnisse eindeutig, dass dieses Modell sich nicht für (russische) Graphemhäufigkeiten eignet: Nur drei der Stichproben kommen auf einen Wert von $C < 0.02$ bei einem Diskrepanzkoeffizienten von $C = 0.029$ für das Gesamtkorpus.

Damit scheiden beide Modelle aufgrund der Befunden für die weiteren Betrachtungen aus, in deren Verlauf wir uns als nächstes der geometrischen und der Good-Verteilung zuwenden wollen. Tab. 5a/b zeigt die Ergebnisse der Anpassungen im Detail.

Wie die Ergebnisse der Tab. 5a/b zeigen, eignen sich auch die geometrische und die Good-Verteilung nicht zur Modellierung russischer Graphemhäufigkeiten: Die Werte des Diskrepanzkoeffizienten C liegen für die einzelnen Stichproben bei der geometrischen Verteilung im Intervall von $0.337 \geq C \geq 0.0167$, für das Gesamtkorpus bei $C = 0.0211$, und nur fünf Stichproben kommen auf einen Wert von $C < 0.02$. Ähnlich schlecht sind die Befunde für die Good-Verteilung: Nur in einem Fall ist der Wert von $C < 0.02$, für das gesamte Korpus beträgt er $C = 0.0211$.

Tab. 4a/b:

		rechts-gestutzt Zeta, R=32			Zipf-Mandelbrot (a,b) n=32			
Nr.	**Kürzel**	**a**	$\chi^2_{FG\ 29}$	**C**	**a**	**b**	$\chi^2_{FG=28}$	**C**
1	ASP-EO 1	0,6659	2001,81	0,1265	2,2784	12,8065	721,25	0,0456
2	ASP-EO 2	0,6828	1487,25	0,1288	2,7371	16,5558	456,80	0,0396
3	ASP-EO 3	0,6725	1693,67	0,1246	2,7403	17,1914	551,15	0,0405
4	ASP-EO 4	0,6714	1530,24	0,1227	1,4136	5,0791	794,37	0,0637
5	ASP-EO 5	0,6703	1580,15	0,1315	1,7478	7,8099	681,15	0,0567
6	ASP-EO 6	0,6728	1632,08	0,1281	1,3574	4,4606	842,83	0,0661
7	ASP-EO 7	0,6722	1920,19	0,1265	1,6214	6,6847	847,98	0,0559
8	ASP-EO 8	0,6886	1951,23	0,1230	3,0305	19,4182	587,00	0,0370
9	ASP-EO 1-2	0,6726	3481,37	0,1272	2,4360	14,0741	1177,47	0,0430
10	ASP-EO 1-3	0,6725	5128,28	0,1252	2,2610	12,5103	1800,41	0,0439
11	ASP-EO 1-4	0,6721	6622,94	0,1239	1,6966	7,4194	2859,07	0,0535
12	ASP-EO 1-5	0,6710	8179,52	0,1249	1,3976	4,8765	4121,98	0,0630
13	ASP-EO 1-6	0,7172	8257,16	0,1056	1,7176	7,5936	4196,80	0,0537
14	ASP-EO 1-7	0,6714	11690,57	0,1252	3,0800	20,3454	3515,08	0,0376
15	ASP-EO 1-8	0,6737	13646,19	0,1249	1,7081	7,4547	5825,38	0,0533
16	LNT-AK	0,7238	158702,58	0,1187	2,1424	9,9287	50451,88	0,0377
17	LNT-OT	0,6991	13438,62	0,1179	6,5497	52,2365	2852,85	0,0250
18	FMD-PR	0,7113	102121,20	0,1219	1,7772	7,1180	40547,22	0,0484
19	FMD-ZA	0,7119	20745,41	0,1102	1,5282	5,2679	9265,88	0,0492
20	APČ-ČA.	0,7073	18228,33	0,1251	12,0000	118,9459	3586,48	0,0246
21	APČ-DJ.	0,7129	7226,2866	0,1187	2,6542	14,8642	1813,34	0,0298
22	MG-MA.	0,7046	50716,0098	0,1171	2,3162	11,9652	17278,67	0,0399
23	MG-NA	0,6982	7563,71	0,0995	7,6666	65,0316	2028,44	0,0267
24	UR	0,7330	1325,26	0,1644	12,0000	102,3936	155,05	0,0192
25	IN	0,7098	2731,32	0,1460	5,6544	40,2040	478,15	0,0256
26	ASP-EO1+8	0,6766	3938,82	0,1243	2,5939	15,5644	1298,49	0,0410
27	LNT-AK8+1	0,7232	924,12	0,1197	5,1200	36,4961	194,05	0,0251
28	FMD-PR1+6	0,6993	3147,01	0,1067	10,8361	94,8234	504,90	0,0171
29	ASP+LND	0,7200	171135,22	0,1184	12,0000	144,0974	83390,77	0,0577
30	ASP+FMD	0,7068	114782,89	0,1212	5,7090	43,1875	24000,24	0,0253
31	ASP+UR	0,6370	13793,61	0,1176	1,8231	8,3289	5941,93	0,0507
32	LNT+UR	0,7239	159407,53	0,1186	12,0000	167,6190	129723,30	0,0965
33	FMD+IN	0,7108	104658,16	0,1222	2,0404	9,2915	36928,53	0,0431
34	MG+IN	0,6868	11190,42	0,1175	12,0000	117,0180	1833,61	0,0192
35	ASP1-5	0,6901	546,93	0,1265	4,4701	33,2901	147,36	0,0341
36	FMD-2	0,7336	1698,62	0,1174	1,7018	6,1837	691,26	0,0478
37	LNT-4	0,7265	876,40	0,1227	2,9291	16,4118	202,65	0,0284
38	GK	0,7121	1031004,65	0,1185	3,2605	20,1769	252634,64	0,0290

Tab. 5a/b:

		rechts-gestutzt geometrisch, $R = 32$			**Good-1 a, p**			
Nr.	**Text**	**q**	$\chi^2_{FG=29}$	**C**	**a**	**p**	$\chi^2_{FG=29}$	**C**
1	ASP-EO 1	0,9086	443,91	0,0280	0,00000014	0,89976	471,65	0,0298
2	ASP-EO 2	0,9064	313,89	0,0272	0,00000030	0,89811	328,77	0,0285
3	ASP-EO 3	0,9083	399,39	0,0294	0,00000008	0,89966	422,38	0,0311
4	ASP-EO 4	0,9087	420,24	0,0337	0,74387629	0,98000	1716,30	0,1376
5	ASP-EO 5	0,9078	354,32	0,0295	0,00000002	0,89914	373,88	0,0311
6	ASP-EO 6	0,9072	337,20	0,0265	0,00000001	0,89867	356,44	0,0280
7	ASP-EO 7	0,9076	394,38	0,0260	0,74717027	0,98000	2115,18	0,1393
8	ASP-EO 8	0,9064	463,37	0,0292	0,00003339	0,89819	483,07	0,0305
9	ASP-EO 1-2	0,9077	752,14	0,0275	0,00000025	0,89911	795,07	0,0290
10	ASP-EO 1-3	0,9080	1122,48	0,0274	0,00000000	0,89936	1189,19	0,0290
11	ASP-EO 1-4	0,9082	1515,85	0,0284	0,00000018	0,89950	1606,65	0,0301
12	ASP-EO 1-5	0,9083	1878,70	0,0287	0,00000001	0,89956	1988,54	0,0304
13	ASP-EO 1-6	0,9081	2195,71	0,0281	0,00000031	0,89943	2324,36	0,0297
14	ASP-EO 1-7	0,9080	2563,15	0,0274	0,00000016	0,89931	2718,88	0,0291
15	ASP-EO 1-8	0,9078	3015,62	0,0276	0,00000007	0,89918	3188,81	0,0292
16	LNT-AK	0,9002	28735,24	0,0215	0,80841128	0,98000	181444,61	0,1358
17	LNT-OT	0,9038	2734,34	0,0240	0,80619800	0,98000	16828,51	0,1477
18	FMD-PR	0,9003	17699,49	0,0211	0,77618992	0,98000	108400,81	0,1294
19	FMD-ZA	0,9025	4464,71	0,0237	0,00000951	0,89501	4589,25	0,0244
20	APČ-ČA.	0,9034	3271,84	0,0225	0,77321643	0,98000	19431,64	0,1333
21	APČ-DJ.	0,9027	1200,12	0,0197	0,00000001	0,89528	1240,04	0,0204
22	MG-MA.	0,9028	10826,33	0,0250	0,78065173	0,98000	57155,47	0,1319
23	MG-NA	0,9063	2039,73	0,0268	0,00000000	0,89805	2134,88	0,0281
24	UR	0,8940	134,70	0,0167	0,80235312	0,98000	1309,22	0,1624
25	IN	0,8987	358,76	0,0192	0,78144672	0,98000	2787,69	0,1490
26	ASP-EO1+8	0,9076	904,82	0,0285	0,00000007	0,89137	1388,15	0,0438
27	LNT-AK8+1	0,8997	173,62	0,0225	0,78516143	0,98000	972,70	0,1260
28	FMD-PR1+6	0,9043	561,27	0,0190	0,75336999	0,98000	3298,06	0,1118
29	ASP+LNT	0,9008	30774,89	0,0213	0,78285514	0,98000	181446,72	0,1255
30	LNT+FMD	0,9013	20958,83	0,0221	0,78699029	0,98000	128874,45	0,1361
31	ASP+UR	0,9070	3071,67	0,0262	0,00000002	0,89853	3240,76	0,0276
32	LNT+UR	0,9002	28774,37	0,0214	0,78783364	0,98000	169893,70	0,1264
33	FMD+IN	0,9013	19026,55	0,0222	0,77609868	0,98000	111138,41	0,1297
34	MG+IN	0,9061	1923,35	0,0202	0,77619024	0,98000	13128,46	0,1378
35	ASP1-5	0,9061	132,49	0,0306	0,75891235	0,98000	596,99	0,1381
36	FMD-2	0,8987	357,04	0,0247	0,00000015	0,89200	359,32	0,0248
37	LNT-4	0,8990	123,35	0,0173	0,00000070	0,89217	124,93	0,0175
38	GK	0,9018	183528,82	0,0211	0,77652268	0,98000	1099027,81	0,1264

Damit können wir die ersten vier der sechs von uns betrachteten – eigentlich die in der bisherigen Forschung am häufigsten diskutierten – Verteilungsmodelle mitsamt als ungeeignet für die Modellierung der Ranghäufigkeit russischer Grapheme bezeichnen. Insofern stellt sich die Frage, inwiefern die beiden verbleibenden Verteilungen, die negativ hypergeometrische und die Whitworth-Verteilung, zu besseren Ergebnissen führen.

Wie oben bereits erwähnt wurde, ist die negativ hypergeometrische Verteilung verschiedentlich für die Modellierung von Ranghäufigkeiten verwendet worden. So haben Köhler/Martináková-Rendeková (1998) zeigen können, dass sie sich zur Modellierung der Häufigkeiten von Tonhöhe, Tonstärke und Tonlänge einer Chopin-Étude eignet, und Wimmer/Altmann (2001) bzw. Wimmer/Wimmerová (Ms.) haben in Werken von Bach, Beethoven, Liszt und Chopin die Ranghäufigkeiten, mit denen Töne einer gegebenen Tonhöhe vorkommen, ebenfalls erfolgreich mit der negativ hypergeometrischen Verteilung modelliert. Auch auf sprachliche Einheiten ist sie mitunter angewendet worden, so z.B. von Ziegler (2001) auf Wortklassenhäufigkeiten im Portugiesischen. Allerdings ist sie auf rangierte Graphemhäufigkeiten bislang erst einmal angewendet worden, und zwar von Grzybek (2001) auf der Basis eines Textes von A.S. Puškin („Царь Салтан“). Insofern stellt die vorliegende Untersuchung auch eine auf breiterer Basis fundierte Überprüfung des an dem Einzeltext erhaltenen, mit einem Wert von $C = 0.0082$ ausgezeichneten Anpassungsergebnisses, dar.

Tab. 6 veranschaulicht die Güte der Anpassung, wie sie sich für das Gesamtkorpus ergibt: die sich aufgrund der angeführten Parameter ergebenden theoretischen Häufigkeiten NP_i sind neben den beobachteten Häufigkeiten f_i der 32 russischen rangierten Grapheme dargestellt; Abb. 2 repräsentiert die Ergebnisse in anschaulicher Form.

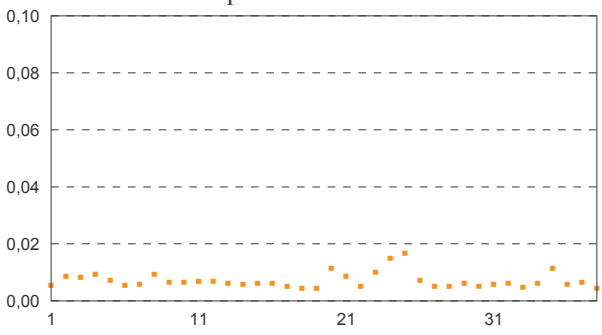
Wie den Daten der Tab. 6 sowie der Abb. 2 zu entnehmen ist, stellt sich in der Tat die negativ hypergeometrische Verteilung als ein ausgezeichnetes Modell dar ($C = 0.0043$). Tab. 7 zeigt die Ergebnisse der Anpassungen für alle einzelnen Stichproben; es bestätigen sich auch hier die hervorragenden Ergebnisse: Der Diskrepanzkoeffizient liegt in allen Stichproben im Intervall von $0.0169 \geq C \geq 0.0043$. In nicht weniger als in 32 der 37 Einzelanalysen ist der Diskrepanzkoeffizient nicht nur $C < 0.02$, sondern sogar $C < 0.01$.

Wie den in Tab. 7 dargestellten Ergebnissen zu entnehmen ist, erweist sich nicht nur der Diskrepanzkoeffizient über alle Einzelstichproben hinweg sowie im gesamten Korpus als überaus stabil; auch die Parameter stellen sich als äußerst konstant dar: Abgesehen von dem ohnehin konstanten Parameter n (der konstant bei $n = 31$, d.h. um eins niedriger als der Inventarumfang liegt), liegen die Werte bei $3.42 \geq K \geq 2.95$ und $0.85 \geq M \geq 0.77$. Abb. 3a/b veranschaulicht die Konstanz der Ergebnisse.

Tab. 6 / Abb. 2: Beobachtete und theoretische Werte (negativ hypergeometrische Verteilung) für das Gesamtkorpus

i	*f*(i)	*NP*(i)	*i*	*f*(i)	*NP*(i)
1	982048	1011265,30	**17**	181684	196234,05
2	763584	774612,97	**18**	163449	177960,39
3	701891	666948,40	**19**	156929	160551,95
4	593949	594729,57	**20**	151944	143961,85
5	563851	538987,63	**21**	146832	128153,72
6	532783	492796,73	**22**	138459	113100,33
7	456610	452879,60	**23**	101994	98782,78
8	423657	417433,13	**24**	93156	85190,15
9	403285	385361,60	**25**	77999	72319,59
10	353818	355950,54	**26**	69870	60177,01
11	295548	328708,97	**27**	54464	48778,55
12	272216	303285,66	**28**	30854	38153,31
13	262459	279421,34	**29**	24421	28348,29
14	248196	256919,77	**30**	22314	19437,96
15	222221	235629,59	**31**	9578	11544,65
16	195629	215432,26	**32**	2257	4891,40
K =3,1441			χ^2 =37465,64		
M =0,7992			FG =28		
n =31			C =0,0043		

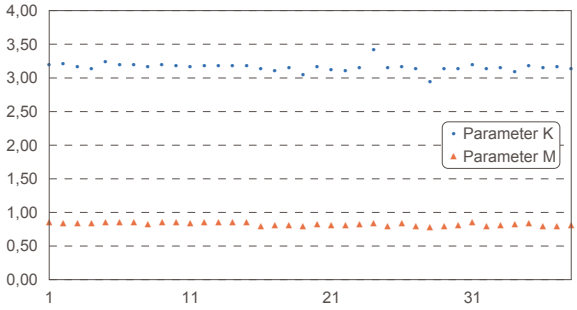
Abb. 3a:
Konstanz des Diskrepanzkoeffizienten *C*



Tab. 7:

Neg. Hypergeometrisch, n=31					
Nr.	**Text**	**K**	**M**	$\chi^2_{FG=28}$	**C**
1	ASP-EO 1	3,1904	0,8472	85,94	0,0054
2	ASP-EO 2	3,2120	0,8394	99,87	0,0087
3	ASP-EO 3	3,1751	0,8405	114,06	0,0084
4	ASP-EO 4	3,1388	0,8306	118,41	0,0095
5	ASP-EO 5	3,2388	0,8531	87,83	0,0073
6	ASP-EO 6	3,2061	0,8450	67,31	0,0053
7	ASP-EO 7	3,2001	0,8445	89,34	0,0059
8	ASP-EO 8	3,1666	0,8250	148,69	0,0094
9	ASP-EO 1-2	3,1974	0,8439	178,68	0,0065
10	ASP-EO 1-3	3,1853	0,8422	269,22	0,0066
11	ASP-EO 1-4	3,1742	0,8397	356,24	0,0067
12	ASP-EO 1-5	3,1816	0,8418	445,28	0,0068
13	ASP-EO 1-6	3,1868	0,8429	480,76	0,0061
14	ASP-EO 1-7	3,1894	0,8434	541,67	0,0058
15	ASP-EO 1-8	3,1869	0,8411	679,85	0,0062
16	LNT-AK	3,1412	0,7893	8231,12	0,0062
17	LNT-OT	3,1084	0,8015	594,66	0,0052
18	FMD-PR	3,1567	0,8005	3839,72	0,0046
19	FMD-ZA	3,0454	0,7818	805,17	0,0043
20	APČ-ČA.	3,1644	0,8137	1691,75	0,0116
21	APČ-DJ.	3,1245	0,8050	533,09	0,0088
22	MG-MA.	3,1065	0,7959	2165,86	0,0050
23	MG-NA	3,1563	0,8259	765,06	0,0101
24	UR	3,4201	0,8269	120,06	0,0149
25	IN	3,1483	0,7927	316,62	0,0169
26	ASP-EO1+8	3,1736	0,8356	229,95	0,0073
27	LNT-AK8+1	3,1401	0,7872	38,47	0,0050
28	FMD-PR1+6	2,9490	0,7707	149,17	0,0051
29	ASP+LNT	3,1400	0,7909	8830,29	0,0061
30	ASP+FMD	3,1465	0,8027	4841,49	0,0051
31	ASP+UR	3,2017	0,8410	686,02	0,0058
32	ASP+UR	3,1422	0,7894	8288,54	0,0062
33	FMD+IN	3,1542	0,8004	4079,82	0,0048
34	MG+IN	3,1014	0,8161	580,80	0,0061
35	ASP1-5	3,1854	0,8282	49,02	0,0113
36	FMD-2	3,1524	0,7816	87,49	0,0060
37	LNT-4	3,1651	0,7910	46,20	0,0065
38	GK	3,1442	0,7992	36999,79	0,0043

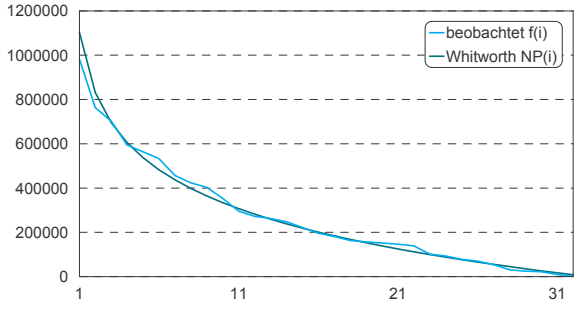
Abb. 3b:
Konstanz der Parameter *K* und *M*



In Anbetracht der Tatsache, dass die negativ hypergeometrische Verteilung drei Parameter aufweist, von denen einer unmittelbar vom Inventarumfang abhängt, ist von besonderem Interesse, dass sich die Whitworth-Verteilung mit ihrem einem Parameter, der ja ausschließlich durch den Inventarumfang vorgegeben ist, als ein ebenso geeignetes Modell für russische Graphemhäufigkeiten erweist. Tab. 8 stellt die Ergebnisse im Detail dar.

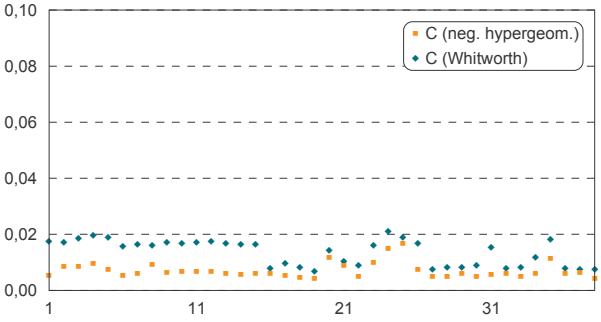
Obwohl die Whitworth-Verteilung nur einen einzigen Parameter (n) hat, der dem Inventarumfang entspricht, ist die Qualität der Ergebnisse durchaus zufriedenstellend: Der Diskrepanzkoeffizient liegt für die einzelnen Stichproben im Intervall $0.212 \geq C \geq 0.0066$; 23 der 37 Einzelstichproben weisen einen Diskrepanzkoeffizienten von $C < 0.02$, 13 gar von $C < 0.01$ auf – nur einer der Texte (der technische Text #28) erreicht nicht das festgelegte Signifikanzniveau von $C < 0.02$. Für das gesamte Korpus beträgt $C = 0.0073$. Abb. 4a veranschaulicht die Güte der Anpassung für das Gesamtkorpus, Abb. 4b stellt die Konstanz dieser Güte über alle 38 Datensätze dar, und zwar im Vergleich mit den Ergebnissen für die negativ hypergeometrische Verteilung (vgl. Abb. 2a).

Abb. 4a:



W h i t w o r t h, R = 32							
Nr.	**Text**	$\chi^2_{FG=30}$	**C**	**Nr.**	**Text**	$\chi^2_{FG=30}$	**C**
1	ASP-EO 1	275,79	0,0174	**20**	APČ-ČA.	2075,78	0,0142
2	ASP-EO 2	196,62	0,0170	**21**	APČ-DJ.	631,01	0,0104
3	ASP-EO 3	252,01	0,0185	**22**	MG-MA.	3880,32	0,0090
4	ASP-EO 4	245,82	0,0197	**23**	MG-NA	1230,06	0,0162
5	ASP-EO 5	227,89	0,0190	**24**	UR	170,95	0,0212
6	ASP-EO 6	201,05	0,0158	**25**	IN	351,32	0,0188
7	ASP-EO 7	246,78	0,0163	**26**	ASP-EO1+8	526,22	0,0166
8	ASP-EO 8	255,72	0,0161	**27**	LNT-AK8+1	57,30	0,0074
9	ASP-EO 1-2	465,81	0,0170	**28**	FMD-PR1+6	236,22	0,0080
10	ASP-EO 1-3	692,18	0,0169	**29**	ASP+LNT	11508,38	0,0080
11	ASP-EO 1-4	908,01	0,0170	**30**	ASP+FMD	8461,09	0,0089
12	ASP-EO 1-5	1142,42	0,0175	**31**	ASP+UR	1798,70	0,0153
13	ASP-EO 1-6	1316,45	0,0168	**32**	LNT+UR	10539,38	0,0078
14	ASP-EO 1-7	1536,60	0,0165	**33**	FMD+IN	7135,16	0,0083
15	ASP-EO 1-8	1784,02	0,0163	**34**	MG+IN	1128,00	0,0118
16	LNT-AK	10464,05	0,0078	**35**	ASP1-5	78,00	0,0180
17	LNT-OT	1094,06	0,0096	**36**	FMD-2	113,15	0,0078
18	FMD-PR	6831,59	0,0082	**37**	LNT-4	54,30	0,0076
19	FMD-ZA	1243,79	0,0066	**38**	GK	63577,01	0,0073

Abb. 4b:
Konstanz des Diskrepanzkoeffizienten *C*



5. Zusammenfassung, Schlussfolgerungen und Perspektiven

Aufgrund der in der vorliegenden Untersuchung angestellten theoretischen Überlegungen und empirischen Befunde ergeben sich eine Reihe von Schlussfolgerungen, aus denen sich weiterführende Perspektiven für zukünftige Forschungen ableiten lassen:

1. Aufgrund der empirischen Befunde gilt es als erstes festzuhalten, dass ganz offenbar auf der Ebene von Graphemanalysen das Problem der Datenhomogenität keine entscheidende Rolle spielt: Die erhaltenen Ergebnisse sind, wie gezeigt werden konnte, konstant, unabhängig davon ob wir es mit vollständigen Texten (verschiedener Definition), mit Textausschnitten, Textkumulationen, Textmischungen, oder einem Korpus zu tun haben.
2. Es gilt festzuhalten, dass vier der üblicherweise im Zusammenhang mit Ranghäufigkeiten von Graphemen diskutierte Verteilungsmodelle – die zeta-Verteilung, die Zipf-Mandelbrot-Verteilung, die geometrische und die Good-Verteilung – sich zumindest für russischen Grapheme nicht eignen; es liegt nahe, dass in dieser Hinsicht eine Reihe von Annahmen auch im Hinblick auf andere Sprachen zu korrigieren sein werden – das jedoch bedarf der empirischen Überprüfung.
3. Als ein gutes, einfaches, und leicht interpretierbares Modell eignet sich die Whitworth-Verteilung, die in der hier dargestellten Form in der Forschung eigentlich noch nicht zum Einsatz gekommen ist; sie zeitigt nicht nur überaus befriedigende Resultate, sondern hat darüber hinaus auch nur einen einzigen Parameter, der dem Inventarumfang gleich ist und insofern direkt interpretiert werden kann.
4. Als ein ausgezeichnetes Modell für die Modellierung rangierter Graphemhäufigkeiten des Russischen eignet sich die negativ hypergeometrische Verteilung, mit der sich hervorragende Anpassungsergebnisse erzielen lassen. Problematisch bei dieser Verteilung ist lediglich, dass sie über nicht weniger als drei Parameter verfügt, von denen nur einer (n) sich direkt interpretieren lässt; die Konstanz der beiden anderen Parameter (K und M) deutet allerdings darauf hin, dass eine Interpretation auch dieser beiden Parameter möglich sein sollte – dafür spricht vor allem die Güte der Whitworth-Verteilung, die ja ein Spezialfall der partial summierten negativen hypergeometrischen Verteilung ist (s.o.), was gegebenenfalls als Indiz dafür zu werten ist, dass die Parameter K und M in zumindest indirekter Abhängigkeit vom Inventarumfang zu sehen sind. Eine Überprüfung dieser Annahme kann jedoch nur in weiteren empirischen Studien mit unterschiedlichem Inventarumfang vorgenommen werden; entsprechende Untersuchungen sind für verschiedene slawische Sprachen

bereits in Arbeit – vgl. erste Ergebnisse von Grzybek/Kelih (2004) zum Slowenischen, und Grzybek/Kelih (2005) zum Slowakischen.

5. Die Ausdehnung der in der vorliegenden Studie durchgeführten Untersuchungen auf weitere Sprachen ist auch aus anderen Gründen notwendig, um zu sehen, inwiefern die hier diskutierten Modelle von über das Russische hinausgehender Relevanz sind. Nicht zuletzt ergeben sich so Einsichten in die graphematischen Strukturen verschiedener (slawischer) Sprachen, incl. historisch-diachronischer Fragen orthographischer Natur.
6. Bei der vertiefenden Interpretation und Ausdehnung der Untersuchungen auf weitere Sprachen wird es von besonderer Bedeutung sein, Querbezüge zur phonologischen Struktur der jeweiligen Sprachen zu kontrollieren – so ist es durchaus möglich, dass sich die hier diskutierten Modelle insbesondere für slawische Sprachen als besonders geeignet erweisen, die eine relativ große (wenn auch unterschiedliche) Nähe zur jeweiligen Phonologie der Sprachen aufweisen.[5]
7. Abgesehen von einer Erweiterung der Untersuchung auf andere (slawische) Sprachen ist eine theoretische Vertiefung der diskutierten Verteilungsmodelle notwendig. Insbesondere wird es notwendig sein, nicht nur weitere empirische Kenngrößen wie Wiederholungsrate und Entropie zu bestimmen, sondern z.B. auch die theoretische Entropie und theoretische Wiederholungsrate der diskutierten Verteilungen zu bestimmen und zu testen – was bislang nur für vereinzelte Verteilungsmodelle geschehen ist (vgl. Zörnig/Altmann 1983, 1984), um zu gesicherten Erkenntnissen zu gelangen (vgl. Grzybek/Kelih/Altmann 2005).

Es stellt sich in der Gesamtzusammenfassung jedenfalls heraus, dass die Untersuchung von Graphemhäufigkeiten weit über das „einfache Zählen“ von Buchstaben hinausgeht und weitreichende Perspektiven für Empirie und Theorie beinhaltet.

[5] Vorläufige Untersuchungen in dieser Richtung zeigen, dass bei der Verfolgung dieser Frage nicht nur streng zwischen phonologischen und phonetischen Häufigkeitsanalysen zu unterscheiden sein wird, sondern dass hier offenbar dem Problem der Datenhomogenität noch mehr Tribut gezollt werden muss als bei den Graphemanalysen – worauf im Grunde genommen schon Peškovskij (1925) aufmerksam gemacht hatte.

Literatur

Altmann, G.; Köhler, R. (1996): „„Language Forces‘ and synergetic modelling of language phenomena“. In: Schmidt, P. (ed.): *Glottometrika 15*. Trier, 62-76.

Altmann, G.; Lehfeldt, W. (1980): *Einführung in die Quantitative Phonologie*. Bochum.

Good, I.J. (1969): „Statistics of Language: Introduction.“ In: Meetham, C.A.; Hudson, R.A. (eds.): *Encyclopaedia of Linguistics, Information, and Control*. Oxford et al., 567-581.

Grzybek, P. (2001): „Kultur-Ökonomie. Zur Häufigkeit text-konstitutiver Elemente.“ In: Weitlaner, W. (Hrsg.): *Sprache – Kultur – Ökonomie*. Wien, 485-509. [= Wiener Slawistischer Almanach, Sonderband 54]

Grzybek, P.; Kelih, E. (2003): „Graphemhäufigkeiten (am Beispiel des Russischen. Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen“. In: *Anzeiger für Slavische Philologie* (XXXI), 131-162.

Grzybek, P.; Kelih, E. (2004): „Rank frequency models for Slovene graphemes.“ In: *Slavistična Revija*. [In print].

Grzybek, P.; Kelih, E. (2005): „Rank frequency models for Slovak graphemes.“ In: Nemcová, E. (ed.): *Philological Studies*. Trnava. [In print].

Grzybek, P.; Kelih, E.; Altmann, G. (2005): „Grapheme Frequencies. Part III: Model characteristics and Criteria.“ [In Vorb.]

Gusein-Zade, S.M. (1988): „O raspredelenii bukv russkogo jazyka po častote vstrečaemosti.“ In: *Problemy peredači informacii* (24/4), 102-107.

Köhler, R.; Martináková-Rendeková, Z. (1998): „A systems theoretical approach to language and music.“ In: Altmann, G.; Koch, W.A. (eds.): *Systems. New Paradigms for the Human Sciences*. Berlin, New York, 514-546.

Krylov, Ju.K. (1982): „Ob odnoj paradigme lingovstatističeskich raspredelenij.“ In: *Lingvostatistika i vyčislitel'naja lingvistika*. Tartu, 80-102.

Martindale, C.; Gusein-Zade, S.M.; McKenzie, D.; Borodovsky, M.Yu. (1996): „Comparison of Equations Describing the Ranked Frequency Distributions of Graphemes and Phonemes“, in: *Journal of Quantitative Linguistics* (3/2), 106-112.

Peškovskij, A.M. (1925): „Desjat' tysjač zvukov. (Opyt zvukovoj charakteristiki russkogo jazyka, kak osnovy dlja eufoničeskich issledovanij).“ In: Dsb., *Metodika rodnogo jazyka, lingvistika, stilistika poėtika. Sbornik statej*. Leningrad/Moskva, 167-191.

Sigurd, B. (1968): „Rank-Frequency Distributions for Phonemes“, in: *Phonetica* (18), 1-15.

Whitworth, W.A. (1901): *Choice and Chance. With One Thousand Exercises*. New York, London 1965.
Wimmer, G.; Altmann, G. (1999): *Thesaurus of univariate discrete probability distributions.* Essen.
Wimmer, G.; Altmann, G. (2000): „On the Generalization of the STER Distribution Applied to Generalized Hypergeometric Parents“, in: *Acta Universitatis Palackianae Olomucensis, Facultas rerum naturalium, Mathematica* (39), 215-247.
Wimmer, G.; Altmann, G. (2001): „Models of Rank-Frequency Distributions in Language and Music.“ In: L. Uhlířová; G. Wimmer; G. Altmann; R. Köhler (eds.): *Text as a Linguistic Paradigm: Festschrift in honour of Luděk Hřebíček*. Trier, 283-294.
Wimmer, G.; Altmann, G. (2003a): „Unified Derivation of Some Linguistic Laws.“ In: *Handbook of Quantitative Linguistics*. [In print]
Wimmer, G.; Altmann, G. (2003a): „Towards a Unified Derivation of Some Linguistic Laws.“ In: Grzybek, P. (ed.): *Word Length Studies and Related Issues*. [In print]
Wimmer, Gejza; Wimmerová, Soňa (Ms.): „Ein musikalisches Rangordnungsgesetz.“
Witten, I.H.; Bell, T.C. (1990): „Source Models of Natural Language Text“, in: *International Journal of Man-Machine Studies* (32), 545-579.
Ziegler, A. (2001): „Word Class Frequencies in Portuguese Press Texts.“ In: L. Uhlířová; G. Wimmer; G. Altmann; R. Köhler (eds.): *Text as a Linguistic Paradigm: Festschrift in honour of Luděk Hřebíček*. Trier, 295-312.
Zörnig, P.; Altmann, G. (1983): „The Repeat Rate of Phoneme Frequencies and the Zipf-Mandelbrot Law.“ In: J. Boy; R. Köhler (eds.): *Glottometrika 5*. Bochum, 205-211.
Zörnig, P.; Altmann, G. (1984): „The Entropy of Phoneme Frequencies and the Zipf-Mandelbrot Law.“ In: J. Boy; R. Köhler (eds.): *Glottometrika 6*. Bochum, 41-47.
Zörnig, P.; Altmann, G. (1995): „Unified representation of Zipf distributions“, in: *Computational Statistics & Data Analysis* (19), 461-473.

Peter Grzybek (Graz): peter.grzybek@uni-graz.at
Emmerich Kelih (Graz): emmerich.kelih@uni-graz.at