Anzeiger für Slavische Philologie
(XXXIII) 2005, 117–140

Graphemhäufigkeiten (Am Beispiel des Russischen)

Teil III: Die Bedeutung des Inventarumfang –

Eine Nebenbemerkung zur Diskussion um das *ë*

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

1. Einleitung

Die vorliegende Studie ist der dritte Teil und die Fortsetzung einer Serie von Untersuchungen zur Vorkommenshäufigkeit russischer Grapheme: Im ersten Teil (Grzybek/Kelih 2003a) ging es zunächst um eine historische Darstellung zur Erforschung von Graphemhäufigkeiten des Russischen inklusive einer Reihe allgemeiner methodologischer Bemerkungen. Der zweite Teil (Grzybek/Kelih/Altmann 2004) war der Frage eines einheitlichen Modells gewidmet; dabei ging es nicht primär um die Häufigkeit einzelner Grapheme, sondern um ein allgemeines Ranghäufigkeitsmodell. Die Leitfrage richtete sich also darauf, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, zum dritthäufigsten, usw. hat – Ziel des Vorgehens war die theoretische Modellierung im Sinne einer mathematischen Formalisierung des Abstands zwischen den jeweiligen Häufigkeiten.[1]

1 Zur Erinnerung sei das Vorgehen kurz zusammengefasst: Überführt man erhobene Ausgangsdaten in eine Rang-Reihenfolge in absteigender Reihenfolge und verbindet dann die Datenpunkte miteinander, ergibt sich charakteristischerweise eine spezifische, monoton fallende Kurve. Bei der Modellierung der Form dieser Kurve geht es darum zu sehen, inwiefern verschiedene Stichproben ein und dieselbe Form aufweisen. Da wir es mit diskreten Häufigkeiten zu tun haben, arbeiten wir allerdings anstelle von Kurven mit diskreten Häufigkeitsmodellen bzw. Wahrscheinlichkeitsfunktionen, was u.a. die Berechnung bestimmter Eigenschaften wie z.B. Entropie, Wiederholungsrate, Momente u.a. ermöglicht. Bei der Erarbeitung dieser Ergebnisse rekurrierten wir auf Grundannahmen der synergetischen Linguistik, insbesondere auf die bei Verteilungsproblemen im allgemeinen vertretene Annahme, dass die Wahrscheinlichkeit einer bestimmten Klasse mit der Ausprägung x oder dem Rang r sich propor-

Als wesentliche Ergebnisse unserer Untersuchung konnten wir die folgenden allgemeinen Punkte festhalten:

a. die Häufigkeiten russischer Grapheme sind nicht zufällig, sondern gesetzmäßig organisiert;
b. sie lassen sich sehr gut durch ein diskretes Wahrscheinlichkeitsmodell beschreiben;
c. alle von uns untersuchten Stichproben folgten ein und dem selben Modell;
d. dieses Modell ist gleichermaßen für vollständige Texte wie für Textausschnitte, Textkumulationen, Textmischungen, und ein vollständiges Textkorpus gültig;
e. das Modell gilt für verschiedene Texttypen und Textsorten (wie etwa prosaische, poetische, technisch-wissenschaftliche u.a.m.).

Als geeignet stellten sich in der Untersuchung 2004 zwei Verteilungsmodelle[2] heraus: 1. die Whitworth-Verteilung, die nur einen Parameter aufweist (nämlich *k* als Inventarumfang), 2. die negative hypergeometrische Verteilung (NHG), die mit außer dem Parameter *k* mit *K* und *M* zweite weitere aufweist. Im Gegensatz zur Whitworth-Verteilung hat sich im Hinblick auch auf andere slawische Sprachen mittlerweile die NHG bestens bewährt – siehe etwa zum Slowenischen Grzybek/Kelih (2003b, 2006b), zum Slowakischen (Grzybek/Kelih/Altmann 2006a,b) oder zum Ukrainischen (Grzybek/Kelih 2005). Aus diesem Grund wollen wir uns im weiteren Verlauf unserer Überlegungen auf dieses Modell – in 1-verschobener Form – beschränken; vgl. Formel (1):

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

---

tional 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 $P_x = g(x)\, P_{x-1}$ aufstellen: Die konkrete Lösung dieser Gleichung hängt natürlich von der jeweiligen Funktion $g(x)$ ab: Dabei gibt man in der Regel einem solchen Modell den Vorzug, das nicht nur auf einen guten Anpassungswert kommt, sondern auch möglichst wenig Parameter aufweist: ein solches Modell ist in der Regel leichter interpretierbar, so dass der Weg von der quantitativen zur qualitativen Analyse beschritten werden kann.

[2] Interessanterweise hat sich in unserer Untersuchung eine Reihe von Modellen als ungeeignet erwiesen, die in der Vergangenheit fast ausschließlich bei Ranghäufigkeitsverteilungen sprachlicher Daten betrachtet wurden; zu diesen ungeeigneten Verteilungen gehörten unter anderem (1) die Zipf-Verteilung, (2) die Zipf-Mandelbrot-Verteilung, (3) die geometrische Verteilung und (4) die Good-Verteilung.

Der hier vorgelegte dritte Teil unserer Untersuchungen zu russischen Graphemen knüpft an eben diese Überlegungen an und präzisiert sie in folgendem entscheidenden Punkt: In der erwähnten Untersuchung von 2004 wurde von einem Inventar von 32 Graphemen ausgegangen; Grund dafür war der Umstand, dass in keinem der Texte der Buchstabe [Ёё] vorkam. Dieser gehört zwar prinzipiell zum normativen Graphembestand des Russischen – der sich damit auf 33 beläuft (s.u.) –, ist in der Praxis allerdings nicht immer konsequent realisiert, sondern durch das einfache [Ee] ersetzt worden,[3] – eben auch in den untersuchten Texten.

[3] Die mehr oder weniger erfolgreiche Geschichte des Buchstabens [Ёё] wird in der Regel auf das Jahr 1783 zurückgeführt, wo bei einer Sitzung der Petersburger Akademie der Wissenschaften durch Fürstin Ekaterina Romanova Daškova die Frage aufgeworfen wurde, ob anstelle der Schreibweise von »йолка« nicht die alternative Schreibweise »ёлка« besser geeignet (ökonomischer) sei. In weiterer Folge festigte sich nach einer entsprechenden Entscheidung durch die Akademie die Schreibung des [Ёё], wobei insbesondere die Werke von N. Karamzin aufgrund ihrer hohen Auflagen einen wesentlichen Beitrag zur Popularisierung des [Ёё] leisteten. Auch zu Zeiten der Sowjetunion wurde der Buchstabe [Ёё] im Alphabet verwendet, wobei seine heute noch gültige normative Verankerung in den Orthographieregeln von 1956 erfolgte (vgl. Pravila 1956: 11):

Bukva ë pišetsja v sledujuščich slučajach:

1. *Kogda neobchodimo predupredit' nevernoe čtenie i ponimanie slova, naprimer: uznaëm v otličie ot oznáem; vsë v otličie ot vse, vëdro v otličie ot vedró; soveršënnyj (pričastie) v otličie ot soveršénnyj (prilagatel'noe).*
2. *Kogda nado ukazat' proiznošenie maloizvestnogo slova, naprimer: reka Olëkma*
3. *V special'nych tekstach: bukvarjach, škol'nych učebnikach russkogo jazyka, učebnikach orfoėpii i t.p., takže v slovarjach dlja ukazanija mesta udarenija i pravil'nogo proiznošenija.*

Die in der alltäglichen Praxis nicht obligatorische Verwendung [Ёё] wurde durch dessen geringe paradigmatische Frequenz begünstigt, so dass immer wieder die Frage diskutiert wurde, ob das russische Alphabet nicht auch ohne diesen Buchstaben auskommen könnte. Gerade in den letzten Jahren jedoch hat das [Ёё] eine ungeahnte Renaissance erlebt: So erschien im Jahre 2000 eine Monographie mit dem Titel *Zwei Jahrhunderte des russischen Buchstaben ë* (Čumakov/Pčelev 2000), die 9500 russische Wörter mit [Ёё] in Form eines Wörterbuchs enthält; 2004 fand eine Konferenz zum Thema „Bukva Ё – neot''emlemaja čast' russkogo jazyka" statt; mittlerweile gibt es Computerprogramme (Makros), die eine nachträgliche Ersetzung fehlender (als „e" geschriebener) Vorkommnisse von [Ёё] ersetzen. Interessanterweise hat auch die *Literaturnaja Gazeta* ab 2004 das [Ёё] wieder programmatisch umgesetzt.– Abgesehen von historischen, kulturgeschichtlichen und nicht zuletzt auch psycholinguistischen Dimensionen wurde die Frage des [Ёё] damit in letzter Zeit bis hin zu einem Problem der nationalen (kulturellen) Identität stilisiert. Die vorliegende Untersuchung versteht

Der Frage der Auswirkung der Berücksichtigung von [Ëë] auf das Graphemsystem in dessen Häufigkeitsstruktur wollen wir in der vorliegenden Studie deshalb detaillierter nachgehen; das soll auf der Basis neuer Texte getan werden, wobei wir (a) im Gegensatz zur Untersuchung von 2004 diesmal tatsächlich homogene Texte (und keine Mischungen, Kumulationen, Segmente, o.a.) verwenden, und (b) gezielt Texte heranziehen, in denen das [Ëë] realisiert wird. Bei der Verfolgung unserer Frage werden wir 30 im folgenden dann als Stichproben bezeichnete Texte in zwei Realisationsformen untersuchen, die wir als B-32 bzw. B-33 bezeichnen werden: zum einen in ihrer Originalform mit [Ëë] (B-33), zum anderen in einer von uns „künstlich" hergestellten Variante, in der alle [Ëë] durch [Ee] ersetzt sind (B-32).

Folgende Fragen sind dabei von vorrangigem Interesse:

1. Wie wirkt sich die unterschiedliche Handhabung des [Ëë] empirisch aus? Zur Beantwortung dieser Frage werden wir globale Kenngrößen von Verteilungen berechnen und für die beiden Untersuchungsbedingungen B-32 und B-33 miteinander vergleichen:
   a.) die Entropie
   b.) die Wiederholungsrate (repeat rate).
2. Wie wirken sich die beiden Realisationen auf der Ebene des theoretischen Verteilungsmodells aus, und zwar
   a.) im Hinblick auf die Anpassungsgüte
   b.) die Parameterwerte der Verteilungsmodelle.

In einem letzten Schritt schließlich werden wir die erhaltenen Ergebnisse in Zusammenhang mit den oben bereits erwähnten Untersuchungen zu anderen slawischen Sprachen stellen, um so die Ergebnisse zum Russischen in ein übergreifendes systematisches Gesamtschema einzubauen, das weitreichende Untersuchungsperspektiven auch für andere Sprachen eröffnet.

2. Datenbasis und prozentualer Anteil des [Ëë]

Unserer Untersuchung werden 30 Texte aus jeweils fünf unterschiedlichen Textsorten zugrunde gelegt. Ausschlaggebend für die Auswahl der Texte war – neben der Breite von unterschiedlichen Texten[4] – vor allem die obligatorische Realisierung des [Ëë].

---

sich allerdings dezidiert nicht als ein direkter Beitrag zu diesen Aspekten der Frage, sondern beschränkt sich auf die rein systematischen Auswirkungen auf der Ebene der Graphemhäufigkeiten.

[4] Die Texte sind der im Rahmen des FWF-Projekts P-15485 »Wortlängenhäufigkeiten in Texten slawischer Sprachen« eingerichteten Text-Datenbank entnommen (vgl.: http://www-gewi.uni-graz.at/quanta).

Für jeden Text werden die Graphemhäufigkeiten bestimmt: zuerst in der originalen Realisationsform (B-33), daran anschließend mit denselben Texten, in denen allerdings [Ëë] durch [Ee] ersetzt ist (B-32). Tab. 1 enthält eine Auflistung der analysierten Texte; der Stichprobenumfang *N* ist natürlich für beide Bedingungen unverändert.

Tabelle 1: Untersuchte Textbasis

Nr.	Autor	Text	Kapitel/Akt	N	Anteil [Ëë] in %
1	A.P. Čechov	*Čajka*	1	15932	0.0690
2			2	12425	0.0161
3			3	10507	0.0381
4			4	14022	0.0214
5		*Jubilej*	1	14161	0.0141
6	A.S. Puškin	*Grobovščik*		10632	0.0282
7		*Metel'*		17925	0.0446
8		*Stancionnyj smotritel'*		17567	0.0342
9		*Vystrel'*	1	10418	0.0288
10			2	7786	0.0385
11	N.A. Nekrasov	*Da, naša žižn' …*		1280	0.2344
12		*Moe razočarovan'e*		2157	0.0464
13		*Muza*		1756	0.1139
14		*Ogorodnik*		1878	0.0532
15		*Rodina*		2060	0.0971
16	L.N. Tolstoj	*Privatbriefe*	1	3580	0.0559
17			2	5982	0.0167
18			3	18464	0.0054
19			4	4290	0.0932
20			5	4909	0.0407
21	I.A. Gončarov	*Oblomov*	II,6	12607	0.0159
22			II,8	22597	0.0044
23			III,7	20945	0.0048
24			III,8	5545	0.0180
25			IV,9	24459	0.0041
26	A.S. Puškin	*Evgenij Onegin*	2	11544	0.0606
27			3	11685	0.0685
28			4	12475	0.0240
29			5	12315	0.0893
30			6	12742	0.0392

Ein erster einfacher deskriptiver Befund unserer Untersuchung ist, dass das [Ëë] insgesamt – was im Grunde genommen zu erwarten war – einen verschwindend kleinen Anteil einnimmt: Für die einzelnen Texte liegt der Anteil bei maximal 0.2344% und der minimalste Anteil nimmt gerade einmal 0.0041% ein.

In den weiteren Überlegungen werden wir folgende Bezeichnungen benutzen:

- N – Stichprobenumfang
- k – Inventarumfang
- f_i – absolute Häufigkeit des Buchstaben i
- p_i – relative Häufigkeit des Buchstaben i, d.h. f_i/N
- ld – logarithmus dualis, $\log_2$
- ln – logarithmus naturalis, $\log_e$

3. Auswirkungen: Entropie und Wiederholungsrate (repeat rate) als globale Eigenschaften

Eine Eigenschaft ist dann als global anzusehen, wenn sie sich nicht auf einzelne Entitäten, sondern auf das ganze Ensemble bezieht. Bei Verteilungen sind es alle Momente, Funktionen wie Variationskoeffizient, Schiefe, Exzess (die in der Linguistik eher selten benutzt werden), oder Eigenschaften wie Wiederholungsrate, Entropie u.a., die man recht häufig antrifft. Die Berechnung solcher globalen Eigenschaften eignet sich also zur Beantwortung der Frage, ob es auf der globalen Ebene zu Einwirkungen kommt; dieser Frage wollen wir im Folgenden durch die Berechnung von Entropie und Wiederholungsraten unter beiden Bedingungen nachgehen.

3.1. Entropie und Relative Entropie

Am bekanntesten ist vermutlich die Entropie, die im Rahmen der Informationstheorie seit ihrer Etablierung durch Shannon in der Regel als ein spezifisches Informationsmaß definiert worden ist, nach der Formel

$$H_1 = -\sum_{i=1}^{k} p_i \cdot \mathrm{ld}\ p_i \,. \tag{2}$$

Die Entropie lässt sich als ein Maß der Gleichverteilung verstehen, d.h. je größer die Entropie, desto ähnlicher sind die Vorkommenshäufigkeiten. Es lässt sich leicht zeigen, dass H_1 sich im Intervall <0; ld k> bewegt, d.h. das (theoretische) Minimum von $H_1 = 0$, und das Maximum ist $H_1 = \mathrm{ld}\ k$, wobei k hier den Inventarumfang von 32 oder 33 bezeichnet.

Damit hängt H_l ganz klar vom Inventarumfang k ab. Dieser Umstand wird oft nicht angemessen berücksichtigt, wenn die Entropien zweier Stichproben miteinander verglichen werden. Wenn wir also die Entropiewerte unserer 30 Stichproben für B-32 und B-33 miteinander vergleichen wollen, dann ist von vornherein zu erwarten, dass die Entropie für B-33 größer sein müsste als für B-32 – was den Vergleich überflüssig erscheinen ließe. Insofern liegt es nahe, in Kenntnis des theoretischen Minimums und des theoretischen Maximums H_1 zu relativieren und die relative Entropie zu berechnen:

$$(3)\quad H_{rel} = \frac{-\sum_{i=1}^{k} p_i \cdot \mathrm{ld}\ p_i}{\mathrm{ld}\, k} = \frac{H_1}{H_0}$$

2.3. Wiederholungsrate und Relative Wiederholungsrate

Ein weiteres, in der Linguistik verbreitetes globales Maß ist die Wiederholungsrate R (repeat rate); sie wird auch als Simpsonsches Distanzmaß oder Herfindahlsches Konzentrationsmaß bezeichnet und wurde von Herdan (1962: 36ff., 1966: 271ff.) in die Linguistik eingeführt. Wie man an der Definition sieht, drückt sie auch die Euklidische Entfernung der Daten in ihrer Gesamtheit vom Ursprung aus:

$$(4)\qquad R = \sum_{i=1}^{k} p_i^2 ,$$

Es handelt sich also schlicht und einfach um die Summe der quadrierten Wahrscheinlichkeiten (relativen Häufigkeiten). Die Wiederholungsrate R hängt offensichtlich ebenso wie H_1 mit dem Inventarumfang (k) zusammen, und ist ebenso wie diese als ein Maß für die Gleichverteilung der Einheiten zu interpretieren: je ähnlicher die Häufigkeiten, desto kleiner wird R: wenn alle Einheiten gleich sind, dann ist $R = 1/k$, denn $\Sigma(1/k)^2 = k(1/k)^2 = 1/k$; und wenn eine der Einheiten die Wahrscheinlichkeit 1 hat und alle anderen 0, dann ist $R = 1$. Auch R, das somit im Intervall $< 1/k; 1 >$ liegt, wird folglich normiert, wenn man den Einfluss des Inventarumfangs relativieren will, und zwar zu

$$(5)\qquad R_{rel} = \frac{1-R}{1-1/n} .$$

Für eine andere Art der Relativierung s. McIntosh (1967).

Zusammenhang von Entropie und repeat rate

In der Regel reicht es aus, eines dieser globalen Maße anzugeben, da diese auf spezifische Art und Weise zusammenhängen und ineinander überführt werden können. So hängen Wiederholungsrate R und Entropie H_1 wie folgt zusammen:

$$H_1 \approx \log_2 k - \frac{k\,R - 1}{2\ln(2)} \qquad \text{bzw.} \qquad R \approx \frac{2(\log_2 k - H_1)\ln(2) + 1}{k}$$

Wie zu sehen ist, sind diese Relationen nur asymptotisch. Beide Indizes lassen sich als Spezialfälle verschiedener verallgemeinerter Entropiearten darstellen (vgl. z.B. Esteban/Morales 1995).

2.4. Vergleich von relativer Entropie und Wiederholungsrate

Tab. 2 enthält für alle 30 Stichproben die Werte der Entropie H_1, der relativen Entropie H_{rel}, der Wiederholungsrate R sowie der relativen Wiederholungsrate R_{rel}.

Konzentrieren wir uns zunächst auf die Entropie: Wie der Tab. 2 zu entnehmen ist, unterscheiden sich die Werte von H_1 und H_{rel} für die Bedingungen B-32 und B-33 nur geringfügig voneinander. Dies entspricht allerdings einer systematischen, von einem Gesetz vorausgesagten Verschiebung (vgl. Altmann/Lehfeldt 1980: 158, 172; Zörnig/Altmann 1984). Diese Verschiebung ist deutlich auch der graphischen Darstellung der Abb. 1 zu entnehmen, welche die Werte der relativen Entropie für die 30 untersuchten Texte für beide Bedingungen enthält.[5]

[5] Führt man zum Zwecke des Vergleichs der Entropiewerte für die beiden Bedingungen B-32 und B-33 einen nicht-parametrischen Test in Form des Mann-Whitney-U-Tests durch, stellt sich heraus, dass die Unterschiede für H_1 nicht signifikant sind ($z = -1.04$, $p = 0.30$), während sie für H_{rel} signifikant sind ($z = 4.26$, $p < 0.001$).

Tabelle 2: Entropie, rel. Entropie, Repeat-Rate und rel. Repeat-Rate

	H_1		H_{rel}		R		R_{rel}	
	32	33	32	33	32	33	32	33
1	4.46127	4.46716	0.89225	0.88557	0.05577	0.05564	0.97469	0.97387
2	4.46969	4.47140	0.89394	0.88641	0.05524	0.05521	0.97524	0.97432
3	4.48321	4.48680	0.89664	0.88946	0.05465	0.05457	0.97585	0.97497
4	4.44984	4.45203	0.88997	0.88257	0.05603	0.05599	0.97443	0.97352
5	4.49095	4.49248	0.89819	0.89059	0.05407	0.05404	0.97645	0.97552
6	4.47793	4.48061	0.89559	0.88824	0.05487	0.05482	0.97562	0.97471
7	4.43403	4.43806	0.88681	0.87980	0.05664	0.05656	0.97380	0.97292
8	4.44585	4.44908	0.88917	0.88199	0.05586	0.05580	0.97459	0.97370
9	4.43428	4.43707	0.88686	0.87960	0.05679	0.05674	0.97364	0.97274
10	4.48025	4.48383	0.89605	0.88887	0.05384	0.05377	0.97668	0.97580
11	4.47649	4.49200	0.89530	0.89049	0.05248	0.05208	0.97809	0.97754
12	4.47774	4.48192	0.89555	0.88850	0.05401	0.05393	0.97650	0.97563
13	4.48279	4.49151	0.89656	0.89040	0.05398	0.05379	0.97654	0.97578
14	4.47128	4.47578	0.89426	0.88728	0.05425	0.05417	0.97626	0.97538
15	4.46993	4.47761	0.89399	0.88764	0.05385	0.05369	0.97667	0.97589
16	4.44007	4.44500	0.88801	0.88118	0.05587	0.05577	0.97459	0.97374
17	4.47368	4.47545	0.89474	0.88721	0.05469	0.05466	0.97580	0.97488
18	4.41399	4.41465	0.88280	0.87516	0.05819	0.05818	0.97219	0.97125
19	4.44472	4.45226	0.88894	0.88262	0.05672	0.05655	0.97371	0.97294
20	4.44472	4.47123	0.88894	0.88638	0.05672	0.05614	0.97371	0.97335
21	4.42800	4.42966	0.87781	0.87814	0.05778	0.05776	0.97166	0.97169
22	4.42114	4.42168	0.87645	0.87655	0.05777	0.05776	0.97168	0.97168
23	4.44323	4.44381	0.88865	0.88094	0.05650	0.05649	0.97394	0.97299
24	4.47021	4.47206	0.89404	0.88654	0.05504	0.05501	0.97544	0.97452
25	4.45806	4.45857	0.89161	0.88387	0.05582	0.05581	0.97464	0.97370
26	4.47968	4.48491	0.89594	0.88909	0.05381	0.05370	0.97672	0.97587
27	4.49052	4.49631	0.89810	0.89135	0.05319	0.05307	0.97736	0.97653
28	4.48786	4.49026	0.89757	0.89015	0.05321	0.05317	0.97733	0.97642
29	4.48760	4.49480	0.89752	0.89105	0.05310	0.05294	0.97745	0.97665
30	4.49209	4.49573	0.89842	0.89123	0.05321	0.05314	0.97733	0.97645

onzentrieren wir uns zunächst auf die Entropie: Wie der Tab. 2 zu entnehmen ist, unterscheiden sich die Werte von H_1 und H_{rel} für die Bedingungen B-32 und B-33 nur geringfügig voneinander. Dies entspricht allerdings einer systematischen, von einem Gesetz vorausgesagten Verschiebung (vgl. Altmann/Lehfeldt 1980: 158, 172; Zörnig/Altmann 1984). Diese Verschiebung ist deutlich auch der graphischen Darstellung der Abb. 1 zu entnehmen, welche die Werte der relativen Entropie für die 30 untersuchten Texte für beide Bedingungen enthält.[6]

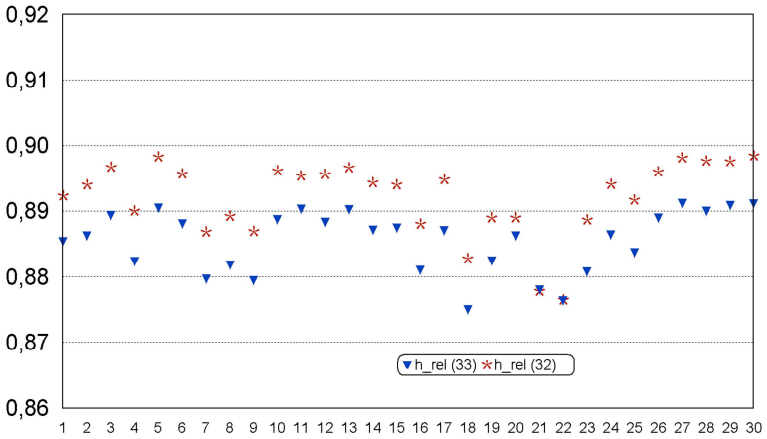

Abb. 1: Relative Entropie (H_{rel}) für B-32 und B-33

Die Tendenz der systematischen Verschiebung bestätigt sich beim Vergleich der Wiederholungsraten für B-32 und B-33, wie die Darstellung in Abb. 2 zeigt, welche die Werte der relativen Wiederholungsrate für die 30 untersuchten Texte für beide Bedingungen enthält.[7]

[6] Führt man zum Zwecke des Vergleichs der Entropiewerte für die beiden Bedingungen B-32 und B-33 einen nicht-parametrischen Test in Form des Mann-Whitney-U-Tests durch, stellt sich heraus, dass die Unterschiede für H_1 nicht signifikant sind ($z = –1.04$, $p = 0.30$), während sie für H_{rel} signifikant sind ($z = 4.26$, $p < 0.001$).

[7] Der Vergleich der Wiederholungsraten für die beiden Bedingungen B-32 und B-33 in Form des Mann-Whitney-*U*-Tests weist die Unterschiede weder für R noch für R_{rel} als signifikant aus ($z = –0.52$, $p = 0.601$ bzw. ($z = –1.88$, $p = 0.06$).

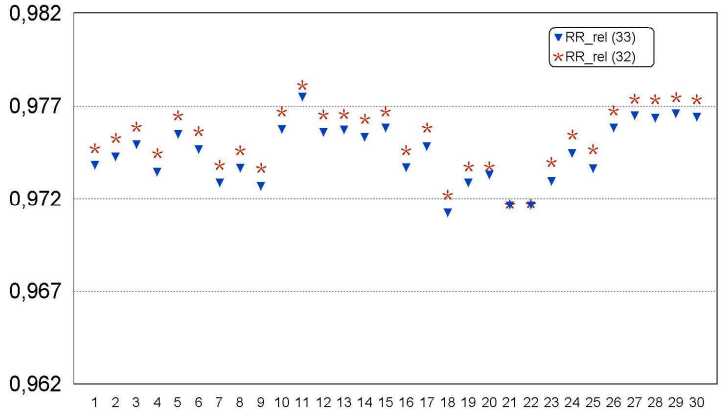

Abb. 2: Relative Wiederholungsraten (R_{rel}) für B-32 und B-33

Eine erste zentrale Schlussfolgerung ist somit, dass eine Entscheidung pro oder contra [Ёё] sich durchaus global auf der Ebene der Häufigkeitsstruktur auswirkt. Abgesehen von gewissen Schwankungen in einzelnen Fällen und der eher schwachen statistischen Signifikanz der Unterschiede kommt es insgesamt zu einer systematischen Verschiebung.

3. Empirische Überprüfung der Verteilungsmodelle

Wie einleitend gesagt, sind bislang zwei Modelle für das Russische passend. Neben der Whitworth-Verteilung ist dies in erster Linie die NHG. Auch für die 30 Texte der vorliegenden Untersuchung bestätigt sich dieser Trend.

Tab. 3 enthält für die Bedingungen B-32 und B-33 jeweils die Parameterwerte für *K* und *M*, den Chi²-Wert der Anpassung der NHG, sowie den hier als Maß der Anpassungsgüte herangezogen Wert des Diskrepanzkoeffizienten *C*.[8]

[8] Die Güte der Anpassung wird üblicherweise mit dem Chi²-Anpassungstest geprüft; in der Linguistik wird bei großen Stichproben statt des Chi²-Werts in der Regel der relativierte Diskrepanzkoeffizient $C = \chi^2/N$ verwendet, wobei ein Wert von $C < 0.02$ als Indiz einer guten, von $C < 0.01$ einer sehr guten Anpassung angesehen wird. Es muss aber bemerkt werden, dass bei extrem großen Stichproben auch *C* seine Schwächen hat.

Tabelle 3: Parameter und Anpassungsergebnisse

Text	**Russisch B-33 = *k***				**Russisch B-32 = *k***			
	K	***M***	$\chi^2_{FG=29}$	***C***	***K***	***M***	$\chi^2_{FG=28}$	***C***
1	3.2459	0.8078	158.91	0.0100	3.0808	0.7969	163.13	0.0102
2	3.2443	0.8119	118.87	0.0096	3.0785	0.8001	110.66	0.0089
3	3.1873	0.8089	121.09	0.0115	3.0377	0.7947	117.04	0.0111
4	3.3553	0.8247	138.16	0.0099	3.1926	0.8094	138.29	0.0099
5	3.2455	0.8270	132.07	0.0093	3.0903	0.8125	122.17	0.0086
6	3.1936	0.8058	78.04	0.0073	3.0456	0.7925	72.15	0.0068
7	3.4102	0.8276	119.96	0.0067	3.2631	0.8160	116.61	0.0065
8	3.3689	0.8280	87.45	0.0050	3.2215	0.8162	74.02	0.0042
9	3.3991	0.8245	54.91	0.0053	3.2324	0.8084	58.23	0.0056
10	3.4575	0.8705	40.24	0.0052	3.2864	0.8529	46.50	0.0060
11	3.3186	0.8638	24.57	0.0192	3.1922	0.8513	25.99	0.0203
12	3.4073	0.8610	23.04	0.0107	3.2517	0.8467	22.25	0.0103
13	3.0837	0.7956	20.81	0.0119	2.9524	0.7832	19.78	0.0113
14	3.1626	0.8087	24.29	0.0129	3.0271	0.7976	21.42	0.0114
15	3.3204	0.8427	22.95	0.0111	3.1769	0.8295	22.82	0.0111
16	3.4917	0.8548	31.99	0.0089	3.3274	0.8391	33.38	0.0093
17	3.1552	0.8013	93.04	0.0156	3.0150	0.7899	81.49	0.0136
18	3.3697	0.8014	130.38	0.0071	3.2048	0.7866	131.93	0.0071
19	3.2145	0.7918	34.41	0.0080	3.0647	0.7792	30.47	0.0071
20	2.9297	0.7510	76.19	0.0155	2.8263	0.7370	69.00	0.0141
21	3.2205	0.7796	87.54	0.0069	3.0669	0.7678	77.19	0.0061
22	3.3499	0.8040	114.78	0.0051	3.1927	0.7909	97.67	0.0043
23	3.2963	0.8079	110.71	0.0053	3.1441	0.7953	84.41	0.0040
24	3.1754	0.8002	40.66	0.0073	3.0337	0.7887	31.72	0.0057
25	3.2726	0.8099	164.32	0.0067	3.1178	0.7964	143.35	0.0059
26	3.3596	0.8524	105.56	0.0091	3.2120	0.8394	99.87	0.0087
27	3.3280	0.8534	102.59	0.0088	3.1862	0.8413	94.75	0.0081
28	3.2853	0.8430	137.85	0.0110	3.1388	0.8306	118.41	0.0095
29	3.3749	0.8640	94.71	0.0077	3.2392	0.8532	86.56	0.0070
30	3.3574	0.8585	77.48	0.0061	3.2061	0.8450	67.31	0.0053

Beim Vergleich zwischen den beiden Bedingungen B-32 und B-33 zeigt sich, dass die Anpassung der NHG unter beiden Bedingungen sehr gut ist: Für B-32 ist in 29 der 30 Stichproben $C < 0.02$, davon in 21 Stichproben $C < 0.01$, und für B-33 ist in allen 30 Stichproben $C < 0.02$, davon ist in 20 Stichproben $C < 0.02$.

Insgesamt stellte sich die NHG als ein ausgezeichnetes Modell[9] dar: Für die einzelnen Stichproben liegt der Diskrepanzkoeffizient im Intervall von $0.0169 \geq C \geq 0.0043$; wenn man alle 30 Texte zu einem Gesamtkorpus zusammenfügt, beträgt $C = 0.0043$.

Damit können wir eine vergleichende Gegenüberstellung vornehmen, welche die prinzipiellen Tendenzen beider Bedingungen (B-32 und B-33) transparent macht und zudem auch eine Vergleichsmöglichkeit mit den Ergebnissen der Studie von 2004 ermöglicht: Für die 30 hier untersuchten Stichproben liegen die Werte des Diskrepanzkoeffizienten für die Bedingung B-32 im Intervall von $0.0203 \geq C \geq 0.0040$, und für B-33 im Intervall von $0.0192 \geq C \geq 0.0050$; damit haben wir es unter beiden Bedingungen mit insgesamt ausgezeichneten Anpassungsergebnissen zu tun, welche die Befunde der 2004er Studie in dieser Hinsicht vollauf bestätigen ($0.0169 \geq C \geq 0.0043$).

4. Parameter *K* und *M* der negativen hypergeometrischen Verteilung

In der Untersuchung von Grzybek/Kelih/Altmann (2004) stellte sich der Diskrepanzkoeffizient *C* als ziemlich stabil dar; das galt auch für die Parameter *K* und *M* der NHG: Abgesehen von dem ohnehin festen Parameter *n* (der konstant bei $n = 31$, d.h. um eins niedriger als der Inventarumfang liegt), wiesen die Parameterwerte mit $3.42 \geq K \geq 2.95$ und $0.85 \geq M \geq 0.77$ eine relativ ausgeprägte Konstanz auf.

Aufgrund der im Anschluss daran durchgeführten Untersuchungen zum Slowenischen, Ukrainischen und Slowakischen[10] ergaben sich allerdings weiterführende Hinweise im Hinblick auf eine mögliche Interpretation der beiden

[9] Die bereits angesprochene Whitworth-Verteilung ist auch für die 30 Stichproben unter den beiden Bedingungen (B-32 und B-33) als akzeptables Modell zu betrachten: Bei B-32 bewegen sich die Werte im Intervall von $0.0366 \geq C \geq 0.0070$ bzw. bei B-33 von $0.0343 \geq C \geq 0.0104$. Wie in Untersuchungen zu anderen slawischen Sprachen (s.u.) allerdings gezeigt werden konnte, eignet sich diese Verteilung bislang nur für das Russische und erlaubt somit keine verallgemeinernden Interpretationen.

[10] Das Slowenische und Slowakische sind in diesem Zusammenhang von besonderer Relevanz, weil sie mit einem Inventarumfang von $N = 25$ (Slowenisch) bzw. $N = 46$ (Slowakisch, mit den Digraphen DZ, DŽ, und CH als eigenen Graphemen) innerhalb der slawischen Sprachen den minimalen und maximalen Inventarumfang repräsentieren.

Parameter *K* und *M* der NHG (ebenso wie vermutlich auch die Parameter anderer Verteilungsmodelle). Demnach wäre es durchaus möglich, dass diese sich direkt oder indirekt auf den Inventarumfang *k* zurückführen lassen (was einer vollständigen Interpretation der Parameter gleichkäme).

Ein erster Schritt, das Verhalten der Parameter *K* und *M* zu präzisieren, ergibt sich somit aus einem Vergleich der Ergebnisse zum Slowenischen[11], Slowakischen, Ukrainischen[12] und zum Russischen aus der 2004er Studie, wobei das Russische im Hinblick auf den Inventarumfang ungefähr in der Mitte des Spektrums liegt.

Wie ein Shapiro-Wilk-Test auf Normalität zeigt, sind die Werte sowohl von *K* als auch von *M* für die beiden Bedingungen B-32 und B-33 jeweils normalverteilt (für *K* ist $p = 0.24$ bzw. $p = 0.20$, für *M* ist $p = 0.12$ bzw. $p = 0.16$). Dieser Umstand erlaubt es uns, neben den Mittelwerten von *K* und *M* auch die Unter- und Obergrenzen der 95%-Konfidenzintervalle zu betrachten, wie sie in Tab. 4 enthalten sind.

Tabelle 4: Inventarumfang und Parameter-Werte

	k	K	K_u	K_o	⁂	M_u	M_o
Slowenisch	25	2.96	2.91	3.00	0.8351	0.8263	0.8439
Russisch (2004)	32	3.16	3.14	3.19	0.8186	0.8105	0.8267
Ukrainisch	33	2.96	2.92	3.01	0.8203	0.8082	0.8324
Slowakisch	46	4.31	4.23	4.40	0.8430	0.8276	0.8584

Es ist leicht zu sehen, dass die Parameterwerte für *K* tendenziell mit zunehmendem *k* ansteigen, wohingegen die Werte für *M* dieser Tendenz so nicht zu folgen scheinen. Dies würde die Beobachtungen und Vermutungen von Grzybek/Kelih 2006 (2006a) stärken, die im Hinblick auf die Analysen der angeführten Sprachen zwei verschiedenen Perspektiven aufgezeigt haben:

1. Die erste Perspektive ist genuin zwischen-sprachlich (sprachübergreifend) und präjudiziert eine direkte Relevanz der Inventargröße *k* für den Parameterwert von *K*; um die Tendenz dieses Parameterverhaltens zu beschreiben, scheint es sinnvoll, sich auf die Mittelwerte von *K* für eine gegebene Sprache zu konzentrieren ($\overline{K}$), so wie dies auch in Tab. 4 der Fall ist.

[11] Wir beziehen uns hier und im folgenden auf die Daten aus Grzybek/Kelih (2006).

[12] Im Hinblick auf das Ukrainische gilt zu bemerken, dass dieses in der ersten Variante ohne Berücksichtigung des Apostroph als eigenständiges Graphem analysiert wurde (Grzybek/Kelih 2005), was den Inventarumfang von $k = 33$ erklärt.

2. Die zweite Perspektive konzentriert sich auf Prozesse innerhalb einer gegebenen Sprache; aufgrund ihrer Beobachtungen schlagen Grzybek/Kelih (2006a) vor, dass *M* zwar nicht wie *K* direkt vom Inventarumfang *k* abhängt (was sich ja nur im zwischensprachlichen Vergleich feststellen lässt), dass aber innerhalb einer gegebenen Sprache der Parameterwert M von *K* abhängt; entsprechend muss *K* dann innerhalb der jeweiligen Sprache (K_i) untersucht werden.

Vor dem Hintergrund dieser Überlegungen ergibt sich nunmehr im Falle des Russischen, die Möglichkeit das Parameterverhalten von *K* und *M* in dieser Hinsicht zu untersuchen und zunächst die Ergebnisse für die Bedingungen B-32 und B-33 vergleichend gegenüberzustellen, um die Ergebnisse sodann abschließend wieder in den Kontext der anderen Sprachen einzubinden.

Für unsere russischen Daten ergeben sich aus diesen allgemeinen Überlegungen somit zwei Hypothesen für die Parameterwerte *K* und *M* unter den beiden Bedingungen B-32 und B-33:

a. aufgrund der unterschiedlichen Inventargrößen sollten die Parameterwerte von *K* für die Bedingung B-33 signifikant größer sein als für B-32; die Parameterwerte für *M* sollten sich nicht signifikant in Abhängigkeit von B-32 vs. B-33 unterscheiden;
b. *M* sollte sowohl für B-32 also auch für B-33 linear von *K* abhängen;
c. $\overline{K}$ sollte sich für B-32 und B-33 systematisch in ein für alle bislang untersuchten Sprachen systematisch strukturiertes Gesamtschema einfügen; zu erwarten wäre eine lineare Regression, innerhalb derer sich das Russische sowohl für B-32 als auch für B-33 einfügt in die bisherigen Ergebnisse zum Slowenischen, Ukrainischen und Slowakischen.

4.1.

Der Mittelwert beträgt für K_{32} = 3.1368 (s = 0.1096); er liegt erkennbar unter demjenigen für K_{33} = 3.2860 (s = 0.1182). Wenn man zum Zwecke des Tests auf Signifikanz dieses Unterschieds der Einfachheit halber den Parameter *K* als eine Zufallsvariable mit einem endlichen Mittelwert und endlicher Varianz betrachtet, dann ergibt ein in Form des *t*-Tests durchgeführter Mittelwertvergleich der Parameterwerte unter Berücksichtigung der unterschiedlichen Inventargröße (k = 32 vs. k = 33), dass der Parameter *K* für k = 33 signifikant größer ist als für k = 32 ($t_{FG=58}$ = 5.07; $p < 0.001$). Die Parameterwerte für M_{32} = 0.8096 (s = 0.0284) und M_{33} = 0.8227 (s = 0.0288) hingegen unterscheiden sich für beide Bedingungen nicht ($t_{FG=58}$ = 1.776; $p < 0.08$).

Dieses Ergebnis zeigt sich deutlich auch an den in Abb. 3 dargestellten Fehlerbalkendiagrammen: Während sich die Konfidenzintervalle für K_{32} und K_{33} nicht überschneiden, ist dies für *M* der Fall.

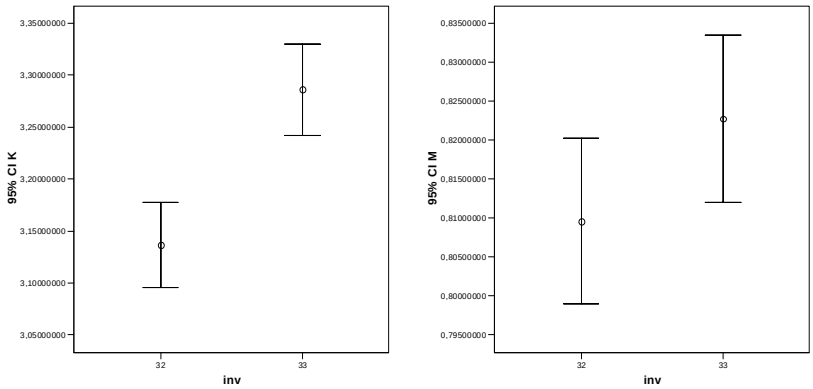

Abb. 3: Konfidenzintervalle und Inventarumfang für B-32 und B33 (links für *K*, rechts für *M*)

4.2.

Es stellt sich heraus, dass sowohl für B-32 als auch für B-33 *M* in der Tat linear von *K* abhängt: für B-32 ist $r = 0.79$ ($p < 0.001$), für B-33 erhalten wir $r = 0.78$ ($p < 0.001$). Damit steht fest, dass für beide Bedingungen jeweils eine lineare Abhängigkeit des Parameters *M* von *K* vorliegt; die diese Abhängigkeit ausdrückende Regressionsgerade folgt der Gleichung $y = a + bx$ (in unserem Fall also $M = a + bK$). Hierbei ist *a* eine Konstante, welche die Höhe der Regressionsgeraden (Schnittpunkt mit der *y*-Achse) angibt, und *b* ist der Regressionskoeffizient, der die Steilheit des Anstiegs bzw. Abfalls der Geraden bestimmt. Für B-32 lautet die Gleichung $y = 0.1653 + 0.2054x$, für B-33 $y = 0.2027 + 0.1887x$. Abb. 4 zeigt deutlich, dass die Regressionsgeraden für die Bedingungen B-32 und B-33 parallel verschoben verlaufen, wohingegen der Anstieg sich nicht erkennbar unterscheidet.

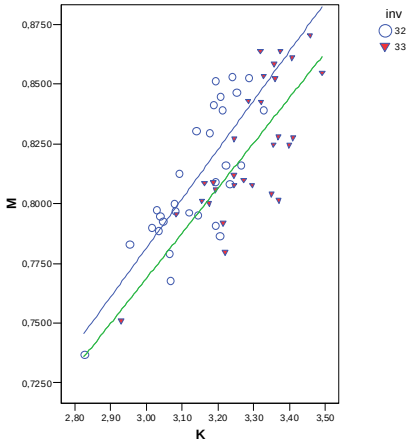

Abb. 4: Systematische Verschiebung der Regressionsgeraden

Diesen Eindruck bestätigt ein Vergleich der beiden Regressionskoeffizienten miteinander und ein statistischer Test auf Unterschiede. Dies geschieht bei linearen Zusammenhängen über die *t*-verteilte Prüfungsgröße

$$(6)\quad t = \frac{\left|b_1 - b_2\right|}{\sqrt{\frac{s^2_{y_1.x_1}\cdot(k_1-2)+s^2_{y_2.x_2}\cdot(k_2-2)}{k_1+k_2-4}\cdot\left(\frac{1}{Q_{x_1}}+\frac{1}{Q_{x_2}}\right)}}$$

bei $FG = k_1 + k_2 - 4$ Freiheitsgraden mit $Q_x = \sum (x - \overline{x})^2$.

Im Ergebnis stellt sich heraus, dass der Unterschied der Regressionskoeffizienten nicht signifikant ist ($t_{FG=56}$ = 0.042; p = 0.99). Wir kommen somit zu der Schlussfolgerung, dass wir es mit einer durch den Inventarumfang *k* bedingten systematischen Verschiebung zu tun haben, was letztendlich der Argumentation von Grzybek/Kelih (2006a) entspricht, der zufolge wohl der Parameter *M* nicht (wie Parameter *K*) direkt vom Inventarumfang *k* abhängt (und deshalb auch nicht von zwischensprachlicher Relevanz ist); vielmehr gibt es zusätzliche Evidenz für die Annahme, dass *M* innerhalb einer gegebenen Sprache von *K* abhängt (und damit von innersprachlicher Relevanz ist). Damit sollten sich die Ergebnisse zum Russischen unter den Bedingungen B-32 und B-33 dann auch in das allgemeine Schema der slawischen Sprachen einfügen.

4.3. Die Parameterwerte *K* und *M* im Gesamtschema

Wir können an dieser Stelle auf die Daten der Tab. 4 rekurrieren; dabei können wir uns zum Zwecke der Anschaulichkeit auf das Minimum und Maximum der Grapheminventare (d.h. das Slowenische und Slowakische) beschränken und im Hinblick auf das Russische die Daten getrennt für B-32 und B-33 hinzufügen. Tab. 5 enthält die diesbezüglichen Daten.

Tabelle 5: Inventarumfang und Parameterwerte

	k	$\overline{K}$	K_u	K_o	$\overline{M}$	M_u	M_o
Slowenisch	25	2.96	2.91	3.00	0.8351	0.8263	0.8439
B-32	32	3.14	3.10	3.18	0.7896	0.7990	0.8202
B-33	33	3.29	3.24	3.33	0.8227	0.8120	0.8335
Slowakisch	46	4.31	4.23	4.40	0.8430	0.8276	0.8584

Die Daten in Tab. 5 legen – wie prognostiziert – ein Ansteigen des Parameters *K* mit zunehmender Inventargröße *k* nahe; dies drückt sich deutlich in der Abb. 5 (im oberen Bereich) aus, die für die vier Teiluntersuchungen getrennt die Fehlerbalkendiagramme (d.h. die Mittelwerte inkl. der Unter- und Obergrenzen der 95%-Konfidenzintervalle aufzeigt. Abb. 5 enthält (im unteren Bereich) in gleicher Weise die Daten für den Parameter *M*, der kein derartiges Verhalten zeitigt; eher scheint *M* über die vier Teiluntersuchungen hinweg relativ konstant zu sein.

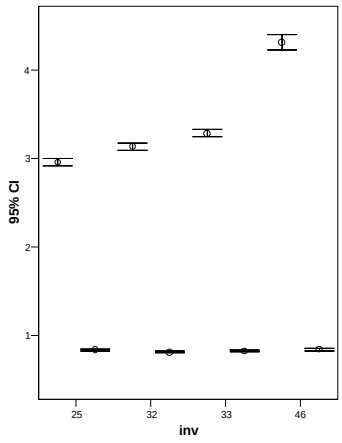

Abb. 5: Inventarumfang und Konfidenzintervalle der Parameter *M* und *K*

Beschränken wir uns zur Verdeutlichung der Trends zunächst in Orientierung an den Werten der Tabl. 5 auf die jeweiligen Mittelwerte von *M* und *K* und berechnen Korrelationen mit dem Inventarumfang *k*. Es stellt sich heraus, dass im Falle von *K* eine signifikante Korrelation vorliegt ($r = 0.98$, $p = 0.025$). Das spricht dafür, dass der Parameter *K* über alle Sprachen hinweg mit dem Inventarumfang korreliert bzw. unmittelbar von diesem abhängt. Eine solche Abhängigkeit besteht zwischen dem Parameter *M* und dem Inventarumfang nicht ($r = 0.42$, $p = 0.59$); das bedeutet, dass der Parameter *M* nicht in gleicher Weise über die verschiedenen Sprachen hinweg unmittelbar vom Inventarumfang *k* abhängt. Dieselbe Tendenz bestätigt sich in noch deutlicherem Maße, wenn man nicht nur die Mittelwerte der vier Teiluntersuchungen betrachtet, sondern alle einzelnen Datenpunkte insgesamt: In diesem Fall ist die Korrelation zwischen *K* und *k* ebenfalls signifikant ($r = 0.94$, $p < 0.001$), diejenige zwischen *M* und *k* hingegen nicht ($r = 0.16$, $p = 0.09$).

Damit liegen die russischen Daten für die beiden Bedingungen B-32 und B-33 im allgemeinen Trend slawischer Sprachen, demzufolge der Parameter *K* im Gegensatz zum Parameter *M* direkt vom Inventarumfang abhängt. Zu klären bleibt abschließend, ob sich an den russischen Daten für die beiden Bedingungen B-32 und B-33 auch die oben aufgestellte Vermutung von Grzybek/Kelih (2006b) bestätigt, der zufolge es einen Zusammenhang zwischen den beiden Parametern *K* und *M* gibt.

Die Berechnung einer Korrelation zwischen *K* und *M* über alle vier Teiluntersuchungen hinweg weist in der Tat den Zusammenhang zwischen den

beiden Parametern als signifikant aus ($r = 0.40$, $p < 0.001$). Die entsprechende Tendenz ist in der Abb. 6 durch die gestrichelte Linie gekennzeichnet. Abb. 6 macht allerdings auch deutlich, dass der Zusammenhang von *K* und *M* innerhalb der einzelnen Sprachen sehr viel deutlicher ausgeprägt ist als über die verschiedenen Sprachen hinweg.[13]

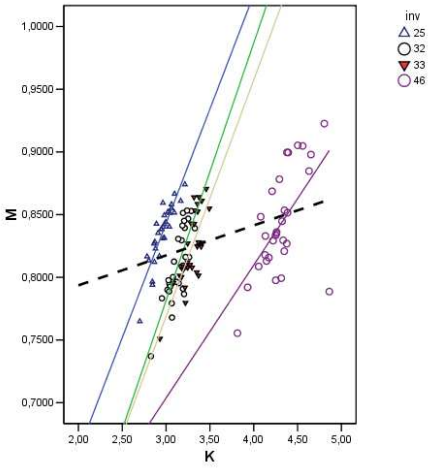

Abb. 6: Zusammenhang der Parameter *K* und *M*

Ersparen wir uns an dieser Stelle die Berechnung der einzelnen Regressionsgeraden und -koeffizienten, wie dies oben im Hinblick auf die beiden russischen Bedingungen B-32 und B-33 bereits getan wurde. Statt dessen sei im Hinblick auf zukünftige und verallgemeinernde Untersuchungen abschließend ein Schema funktioneller Abhängigkeiten skizziert, welches die Zusammenhänge zwischen den Parametern *K* und *M* und dem jeweiligen Inventarumfang *k* verdeutlicht, und welches die im vorliegenden Text beobachteten Befunde zu russischen Graphemhäufigkeiten in die bisherigen Untersuchungen zu Graphemhäufigkeiten in verschiedenen slawischen Sprachen einfügt.

[13] Die entsprechend höheren Werte der Korrelationskoeffizienten betragen im einzelnen: $r = 0.88$ (slowenisch), $r = 0.79$ (Russisch, B-32), $r = 0.77$ (Russisch, B-33), sowie $r = 0.59$ (Slowakisch), wobei in allen Fällen $p < 0.001$).

Dieses allgemeine von Grzybek/Kelih (2006a) entwickelte Schema sieht die folgenden Abhängigkeiten vor: Demnach ist *K* eine sprachübergreifende Funktion von *n*, die durch eine lineare Gleichung ausgedrückt werden kann. Für jede (slawische) Sprache mit einem jeweils bestimmten Inventarumfang *k* gibt es folglich ein relativ stabiles *K*, weshalb es nahe liegt, sich im Hinblick auf eine gegebene Sprache auf den aus verschiedenen Stichproben ergebenden Mittelwert für *K* zu beziehen; die Gleichung lautet demnach

(7) $$\overline{K} = h(k) = u \cdot n + v$$

M hingegen scheint im Gegensatz zu *K* sprachspezifisch relevant zu sein, so dass es sinnvoll ist, hier das M_i innerhalb einer gegebenen Sprache in Betracht zu ziehen, wobei sich diese Funktion nur mit einem Regressionskoeffizienten ohne zusätzliche Konstante darstellen lässt:

(8) $$M_i = g(K_i) = a_i \cdot K_i$$

Der sprachspezifische Regressionskoeffizient a_i ist seinerseits wiederum eine Funktion von *k*, die abermals linearen Charakters ist, d.h.:

(9) $$a_i = f(k) = c \cdot n + d$$

Tab. 6 enthält die für die in der vorliegenden Studien untersuchten Sprachen erhaltenen Werte für a_i.

Tabelle 6: Inventarumfang und Regressionskoeffizienten

k	*K*	*M*	*a*
25	2.9564	0.8351	0.2823
32	3.1368	0.8096	0.2580
33	3.2860	0.8227	0.2503
46	4.3137	0.8430	0.1952

Den Zusammenhang zwischen *a* und dem Inventarumfang, der bei einem Wert von $r = 0.998$ höchst signifikant ist, stellt die abschließend präsentierte Abb. 7 anschaulich dar.

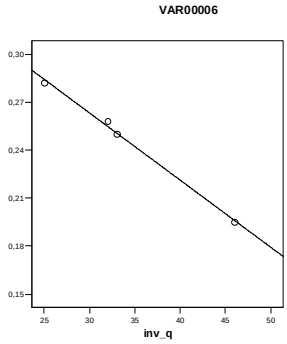

Abb. 7: Zusammenhang zwischen Inventargröße und Regressionskoeffizienten

Eine plausible Erklärung für dieses Modell bestünde darin, dass einerseits über den Parameter *K* eine Anpassung der Häufigkeitsverteilung in Abhängigkeit vom Inventarumfang einer gegebenen Sprache stattfindet, und dass andererseits über den Parameter *M* Schwankungen innerhalb einer gegebenen Sprache bzw. innerhalb einzelner Stichproben ausgeglichen werden. Eine Bestätigung dieses Modells lässt sich nur durch die Analyse weiterer Sprachen erhalten – ungeachtet dessen steht fest, dass sich das Russische mit beiden Bedingungen, B-32 und B-33, nahtlos in dieses Schema einfügt, wenn auch – wie zu sehen – abermals mit systematischer Verschiebung

Resultate und Perspektiven

Die vorgestellte empirische Untersuchung auf der Basis von 30 russischen Stichproben ergibt die folgenden zentralen Resultate:

1. Die unterschiedliche Handhabung des [Ёё] ergibt unter den zwei Bedingungen (B-32 und B-33) auf der Ebene von Entropie und Wiederholungsrate als zwei maßgeblichen globalen Kenngrößen einer Häufigkeitsverteilung, dass die Werte einem gesetzmäßig voraussagbaren Verhalten entsprechen und sich auf systematische Weise verschieben.
2. Als zweiter wichtiger Befund hinsichtlich eines theoretischen Verteilungsmodells lässt sich festhalten, dass sowohl für B-32 als auch für B-33 von einem gemeinsamen einheitlichen Modell ausgegangen werden kann;

hierbei handelt es sich um die negative hypergeometrische Verteilung, deren Parameter sich freilich unter den beiden Bedingungen unterschiedlich verhalten.

3. Die Parameter *K* und *M* der negativen hypergeometrischen Verteilung zeigen ein systematisches Verhalten im Hinblick auf den Inventarumfang (d.h. hinsichtlich B-32 und B-33); dieses Parameterverhalten fügt sich nahtlos in bisherige Befunde zu einem Gesamtbild der Modellierung slawischer Sprachen (Slowenisch, Ukrainisch, Slowakisch) ein.

Eine Absicherung der aufgezeigten Tendenzen und eine detaillierte Untersuchung des Zusammenhangs zwischen den beiden Parametern wird nur durch Berücksichtigung weiterer Sprachen möglich sein, wobei im Idealfall eine qualitative Interpretation der Parameter das Endergebnis sein könnte.

Literatur

Altmann, G.; Köhler, R. (1996): „‚Language Forces' and synergetic modelling of language phenomena". In: Schmidt, P. (Hrsg.): *Glottometrika 15.* Trier: Wissenschaftlicher Verlag, 62–76.

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

Esteban, M.D.; Morales, D. (1995): "A summary of entropy statistics." In: *Kybernetica* (31/4); 337–346.
[vgl.: http://www.cse.msu.edu/~cse902/S03/entropy_measures.pdf]

Grzybek, P.; Kelih, E. (2003a): „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. (2003b): Grapheme Frequencies in Slovene. In: *Slovko 2003*. Bratislava. [in Druck]

Grzybek, P.; Kelih, E. (2005): Graphemhäufigkeiten im Ukrainischen. Teil 1 Ohne Apostroph (’). In: Altmann, G.; Levickij, V.; Perebejnos, V. (Hrsg.) (2005): *Problems of Quantitative Linguistics 2005*. Černivci: Ruta, 159–179.

Grzybek, P.; Kelih, E. (2006a): Towards a General Model of Grapheme Frequencies for Slavic Languages. In: *Slovko 2005.* Bratislava. [in Druck]

Grzybek, P.; Kelih, E. (2006b): „Zur Frage der Stichprobengrößen in Graphem- und Phonemuntersuchungen (am Beispiel slowenischer Grapheme)", in: *Glottometrics*, 13 [in Druck]

Grzybek, P.; Kelih, E.; Altmann, G. (2004): „Graphemhäufigkeiten (am Beispiel des Russischen. Teil II: Modelle der Häufigkeitsverteilung“, in: *Anzeiger für Slavische Philologie* (XXXII), 25–54.
Grzybek, P.; Kelih, E.; Altmann, G. (2006a): Graphemhäufigkeiten im Slowakischen (Teil II: Mit Digraphen). In: *Sprache und Sprachen in Mitteleuropa*. GeSuS, Trnava (2005). [in Druck]
Grzybek, P.; Kelih, E.; Altmann, G. (2006b): Graphemhäufigkeiten im Slowakischen (Teil I: Ohne Digraphen). In Nemcová, E. (Hrsg.): *Philologia actualis slovaca*. UCM, Trnava. [in Druck]
Herdan, G. (1962): *The calculus of linguistic observations.* The Hague: Mouton.
Herdan, G. (1966): *The advanced theory of language as choice and chance.* Berlin: Springer.
McIntosh, R.P. (1967): “An index of diversity and the relation of certain concepts to diversity.” In: *Ecology* (48), 392–404.
Pčelov, E.V.; Čumakov, V.T. (2000): *Dva veka russkoj bukvy ë.* Moskva: Narodnoe obrazovanie.
Pravila (1956): *Pravila russkoj orfografii i punktuacii.* Moskva: Izdatel’stvo Ministerstva Prosveščenija.
Zörnig, P.; Altmann, G. (1984): „The Entropy of Phoneme Frequencies and the Zipf-Mandelbrot Law.“ In: Boy, J.; Köhler, R. (Hrsg.): *Glottometrika 6.* Bochum: Brockmeyer, 41–47.

Peter Grzybek (Graz): peter.grzybek@uni-graz.at