Graphemhäufigkeiten im Slowakischen

(Teil II: Mit Digraphen)

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

1. Graphemhäufigkeiten und Modelle ihrer theoretischen Beschreibung

Die vorliegende Studie ist Teil einer Serie von Untersuchungen zur Vorkommenshäufigkeit von Graphemen in verschiedenen (zunächst einmal vornehmlich slawischen) Sprachen im Allgemeinen, und zur Frequenz slowakischer Grapheme im Besonderen. Den Auftakt zur gesamten Untersuchungsserie stellt eine historische Darstellung zur Erforschung von Graphemhäufigkeiten des Russischen dar (Grzybek/Kelih 2003). Nachdem schon dieser historischen Darstellung eingangs eine Reihe allgemeiner methodologischer Bemerkungen zur Untersuchung von Graphemhäufigkeit vorangestellt worden waren, führten Grzybek/Kelih (2003) und Grzybek/Kelih/Altmann (2005b) eigene Untersuchungen zur Vorkommenshäufigkeit russischer bzw. slowakischer Grapheme durch.

Während in der Vergangenheit das Interesse bei der Untersuchung von Graphemhäufigkeit in der Regel eher auf die relative Häufigkeit des Vorkommens einzelner Grapheme ausgerichtet war, wurde von Grzybek/Kelih/Altmann (2005b) die Frage nach einem allgemeinen Häufigkeitsmodell gestellt. Dabei geht es darum, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, zum dritthäufigsten, usw. hat. Das Interesse richtet sich in diesem Fall somit nicht auf die Häufigkeit spezifischer Grapheme; vielmehr wird die Frage gestellt, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, zum dritthäufigsten, usw. hat. In den Vordergrund rückt damit eine Rang-Häufigkeitsverteilung, das Ziel der theoretischen Modellierung ist 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.

Von den Ergebnissen der o.a. Studie zum Russischen ausgehend, haben Grzybek/Kelih/Altmann (2005b) dann auch im Hinblick auf die Vorkommenshäufigkeit slowakischer Grapheme erstmals die Frage nach einem theoretischen Verteilungsmodell gestellt. Dabei wurden Digraphen konsequent nicht als eigene Einheiten gerechnet, so dass der Inventarumfang entsprechend der heute gültigen slowakischen Orthographie $n = 43$ betrug. In der hier vorliegenden Untersuchung sollen die drei Graphemkombinationen DZ, DŽ, und CH hingegen als eigenständige, zum Systembestand zu zählende digraphische Einheiten angesehen werden, so dass es sich um ein Inventar von 46 Graphemen handelt.

Allein dieser Umstand wirft die Frage nach der Vergleichbarkeit der Untersuchung auf. Die entscheidende Frage ist dabei, ob sich das Modell, das sich in der Studie von Grzybek/Kelih/Altmann (2004) bei einer Inventargröße von 43 Graphemen für das Slowakische als passendes theoretisches Modell herausstellte auch bei Berücksichtigung der drei Digraphen und der damit veränderten Inventargröße als geeignet erweist.

Bei der Durchführung dieser weiterführenden Vergleichsuntersuchung wäre es natürlich möglich, sich auf dasjenige Modell zu beschränken, das sich bei einer Inventargröße von $n = 43$ als passend herausstellte. Damit wäre jedoch eine Einschränkung verbunden, welche die mögliche Eignung anderer Modelle im Vorfeld ausschließen würde. Aus diesem Grund soll in den folgenden Analysen alle in den einschlägigen Diskussionen bislang ins Spiel gebrachten Modelle in gleicher Weise geprüft werden, wie das im Falle der Inventargröße von $n = 43$ war.

Um die mathematische Herleitung der einzelnen Modelle muss es an dieser Stelle nicht im Detail gehen; diese sind ausführlich in Grzybek/Kelih/Altmann (2004) dargestellt. Im einzelnen handelt es sich um die folgenden Verteilungsmodelle[1], die auch in der hier vorliegenden Studie auf ihre Adäquatheit hin geprüft werden sollen (s.u.):

(1) Zipf- / Zipf-Mandelbrot-Verteilung
(2) Zeta-Verteilung
(3) Geometrische Verteilung

[1] Da Graphemsysteme nur eine relativ begrenzte Anzahl unterschiedlicher Klassen aufweisen, ist es sinnvoll, diejenigen Verteilungen, deren Definitionsbereich nicht von *1…n* (sondern bis unendlich) geht, auf der rechten Seite zu stutzen.

(4) Good-Verteilung
(5) Whitworth-Verteilung
(6) Negative hypergeometrische Verteilung

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.

3. Empirische Überprüfung der Modelle

3.1. Text- und Datenbasis

Bei der empirischen Überprüfung der oben dargestellten Modelle an slowakischen Graphemen soll, wie oben bereits angesprochen wurde, dasselbe Textmaterial verwendet werden wie in der Untersuchung von Grzybek/Kelih/Altmann (2004), in der besonderer Wert darauf gelegt wurde, die Datenhomogenität systematisch zu kontrollieren: Zwar sind auf der Ebene der Grapheme nicht unbedingt durch die Verletzung der Datenhomogenität bedingte Inkonsistenzen zu erwarten, doch soll(te) eine entsprechend systematische Kontrolle dieses Faktors auf jeden Fall gewährleistet sein.

Während allerdings in der genannten Untersuchung zum Russischen von Grzybek/Kelih (2003) zu diesem Zweck systematisch und kontrolliert nicht nur vollständige Texte, sondern auch Textausschnitte, Textkumulationen, Textmischungen und ein vollständiges Gesamtkorpus bearbeitet wurden, handelt es sich in der hier vorliegenden Untersuchung ausschließlich 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.

Tab. 1 stellt eine Übersicht über die Texte dar. Bei den ersten zehn Texten handelt es sich um jeweils einzelne Kapitel aus zeitgenössischen Romanen von Rudolf Sloboda und Vincent Sikula, Texte 11-15 sind ausgewählte Kapitel aus slowakischen (geisteswissenschaftlichen) Diplomarbeiten, Texte

16-20 sind journalistische Kommentare, Texte 21-25 Kunstmärchen, und Texte 26-30 schließlich Fachtexte.
In der Tabelle findet sich im Anschluss an die Nummer des Textes – auf die auch in den einzelnen Analysen Bezug genommen werden wird – eine Angabe zum Autor bzw. zur Quelle des Textes, sodann die Bezeichnung des Textes, der Textstatus, und schließlich der Umfang des Textes (in der Anzahl der Buchstabenvorkommnisse).[2]

Tab. 1: Text- und Datenbasis

Nr.	Autor	Text	Kapitel	N
1	Rudolf Sloboda	*Pamäti*	1	6939
2		*Pamäti*	2	16548
3		*Pamäti*	3	4108
4		*Pamäti*	4	3373
5		*Pamäti*	5	8469
6	Vincent Sikula	*Veterná ružica*	1	7026
7		*Veterná ružica*	2	23691
8		*Veterná ružica*	3	10390
9		*Veterná ružica*	4	13669
10		*Veterná ružica*	5	5303
11	anonym	Diplomarbeit l		2482
12		Diplomarbeit 2		5943
13		Diplomarbeit 3		5594
14		Diplomarbeit 4		3802
15		Diplomarbeit 5		5155
16	anonym	Kommentar 1		1606
17		Kommentar 2		3517
18		Kommentar 3		1470
19		Kommentar 4		1607
20		Kommentar 5		2461
21		Kunstmärchen 1		562

[2] Die Texte wurden der im Aufbau befindlichen Text-Datenbank slowakischer Texte entnommen; diese Text-Datenbank wird in einer österreichisch-slowakischen Kooperation zwischen den Universitäten Graz (Institut für Slawistik) und Trnava (Lehrstuhl für Slowakistik) im Rahmen der »Aktion Österreich-Slowakei« (Projekt 43s9) aufgebaut, unter finanzieller Unterstützung durch den Österreichischen Akademischen Austauschdienst (OEAD) und die Slovenská akademická informačná agentúra (SAIA).

22		Kunstmärchen 2	443
23		Kunstmärchen 3	632
24		Kunstmärchen 4	1214
25		Kunstmärchen 5	1338
26	anonym	Fachtext 1	1698
27		Fachtext 2	1482
28		Fachtext 3	4196
29		Fachtext 4	1294
30		Fachtext 5	1380

Wie oben bereits gesagt wurde, soll es in der vorliegenden Untersuchung um die individuellen Texte, wie sie in Tab. 1 aufgeschlüsselt sind, gehen. Dennoch können wir das Vorgehen exemplarisch am Gesamtkorpus dieser Texte veranschaulichen: Fügt man die 30 einzelnen Texte zu einem Gesamtkorpus zusammen, so beläuft sich der Umfang dieses Korpus auf N = 147.392 Vorkommnisse (im Vergleich zu 148.940 Graphemen bei einer zugrunde gelegten Inventargröße von n = 43 ohne Digraphen). Tab. 2 gibt die absoluten und relativen Häufigkeiten für die 46 Grapheme bzw. Digraphen wieder.

Tab. 2: **Graphemhäufigkeiten im Gesamtkorpus**

Graphem	**f(i)**	**f_{rel}(i)**	**Graphem**	**f(i)**	**f_{rel}(i)**
a	14194		**m**	5659	
á	2408		**n**	8323	
ä	172		**ň**	297	
b	2676		**o**	13772	
c	1593		**ó**	131	
č	1825		**ô**	270	
d	5103		**p**	4121	
ď	402		**q**	0	
dz	124		**r**	6164	
dž	2		**ŕ**	27	
e	12701		**s**	7099	
é	947		**š**	1685	
f	253		**t**	6562	
g	346		**ť**	1395	

h	3376	**u**	3845
ch	1422	**ú**	1073
I	9285	**v**	6534
í	1465	**w**	10
j	3135	**x**	47
k	5731	**y**	2262
l	6091	**ý**	1294
Í	3	**z**	2660
l'	719	**ž**	1611

Ordnet man die Vorkommenshäufigkeiten der einzelnen Grapheme in absteigender Reihenfolge, so ergibt sich die Ranghäufigkeitsverteilung, die für das Gesamtkorpus in Abb. 1 veranschaulicht ist.

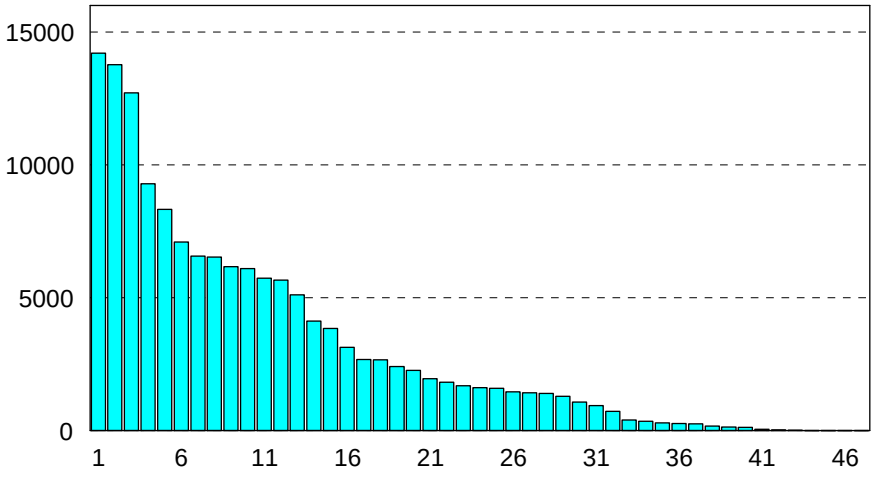

Abb. 1

Ranghäufigkeitsverteilung slowakischer Grapheme

(Gesamtkorpus)

Um die theoretische Modellierung der Ranghäufigkeiten in den 30 slowakischen Texten geht es in den folgenden Analysen.[3]

3.2. Ergebnisse

Schauen wir uns im folgenden die Ergebnisse für die 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 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 Zipf-Mandelbrot-Verteilung an die Daten des Gesamtkorpus an. Tab. 3 enthält neben den Rängen 1 bis 46 die absoluten

[3] Graphisch werden diskrete Häufigkeitsverteilungen wie in Abb. 1 üblicherweise als Balkendiagramme dargestellt; aus Gründen der besseren Anschaulichkeit werden in den unten folgenden Darstellungen Liniendiagramme vorgezogen.

Häufigkeiten *f(i)* in absteigender Reihenfolge. Der Parameter $n = 46$ berechnet sich unmittelbar aus dem Inventarumfang; für die Parameter *a* und *b*, die sich auf verschiedene Arten und Weisen schätzen und dann in iterativen Verfahren optimieren lassen, ergibt sich im gegebenen Fall $a = 12{,}00$ und $b = 117{,}11$. Setzt man diese Werte in die Formel (5) ein, ergeben sich die theoretischen Werte *NP(i)* – vgl. auch Formel (5):

Tab. 3: Ergebnis der Anpassung der rechts-gestutzten Zipf-Mandelbrot-Verteilung an das Gesamtkorpus

i	*f*(i)	*NP*(i)	i	*f*(i)	*NP*(i)
1	14194	13462,65	24	1611	1591,88
2	13772	12167,28	25	1593	1462,54
3	12701	11005,86	26	1465	1344,51
4	9285	9963,58	27	1422	1236,74
5	8323	9027,39	28	1395	1138,25
6	7099	8185,75	29	1294	1048,21
7	6562	7428,46	30	1073	965,84
8	6534	6746,48	31	947	890,43
9	6164	6131,81	32	719	821,35
10	6091	5577,35	33	402	758,05
11	5731	5076,80	34	346	700,00
12	5659	4624,54	35	297	646,73
13	5103	4215,61	36	270	597,82
14	4121	3845,57	37	253	552,90
15	3845	3510,45	38	172	511,61
16	3135	3206,75	39	131	473,64
17	2676	2931,30	40	124	438,71
18	2660	2681,30	41	47	406,55
19	2408	2454,24	42	27	376,92
20	2262	2247,86	43	10	349,62
21	1954	2060,15	44	3	324,45
22	1825	1889,30	45	2	301,24
23	1685	1733,69	46	0	279,81
	$a = 12{,}00$			$\chi^2 = 5140{,}61$	
	$b = 117{,}11$			FG = 43	
	$n = 46$				
	$N = 147392$			C = 0,0349	

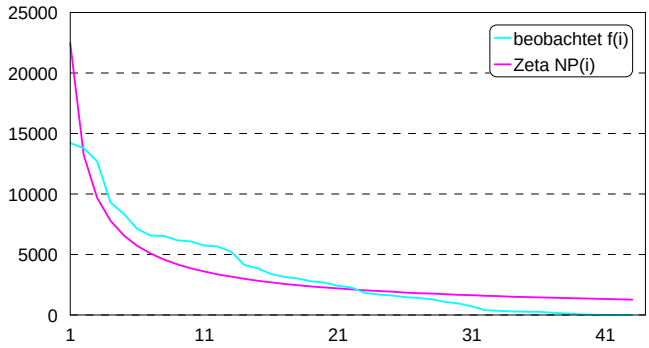

Abb. 2

Anpassung der rechts-gestutzten Zipf-Mandelbrot-Verteilung
(Gesamtkorpus)

Wie den Werten in Tab. 3 und ihrer Veranschaulichung in Abb. 2 zu entnehmen ist, stellt die Zipf-Mandelbrot-Verteilung für die Daten des Gesamtkorpus kein gutes Modell dar; dies drückt sich auch im Wert des Diskrepanzkoeffizienten $C = 0.0349$ klar aus. Das schlechte Ergebnis erklärt es auch, warum mit der rechts-gestutzten Zeta-Verteilung (die ja über einen Parameter weniger verfügt), erst gar keine Anpassung möglich ist.
Diese negative Einschätzung gilt allerdings nicht nur für das gesamte Korpus, sondern für alle anderen Einzelstichproben in gleicher Weise: Tab. 4a/b enthält die Ergebnisse für alle 30 Datensätze. Neben den Ergebnissen für die rechts-gestutzte Zeta-Verteilung (2a/b) enthält Tab. 4a/b auch die Anpassungsergebnisse der 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 slowakischer Grapheme eignen: Zwar gelingt bei den einzelnen Texten – im Gegensatz zum Gesamtkorpus – eine Anpassung der Zeta-Verteilung, doch liegen die Werte des Diskrepanzkoeffizienten für die einzelnen Datensätze im Intervall von $0.26336 \geq C \geq 0.1749$, so dass nicht eine einzige Stichprobe auf einen Wert von $C < 0.02$ kommt. 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 (slowakische) Graphemhäufigkeiten eignet:

Keine einzige der 30 Stichproben kommt auf einen Wert von $C < 0.02$ ($0.1376 \geq C \geq 0.0302$).

Damit scheiden beide Modelle aufgrund der Befunde 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 zwei weitere Modelle, nämlich die geometrische und die Good-Verteilung, nicht zur Modellierung slowakischer Graphemhäufigkeiten: Die Werte des Diskrepanzkoeffizienten C liegen für die einzelnen Stichproben bei der geometrischen Verteilung im Intervall von $0.713 \geq C \geq 0.0208$, für das Gesamtkorpus bei $C = 0.0303$. Ähnlich schlecht sind die Befunde für die Good-Verteilung: Auch hier ist in keinem einzigen Fall der Wert von $C < 0.02$, für das gesamte Korpus beträgt er $C = 0.0301$; interessanterweise tendiert hierbei (wie auch in einer ganzen Reihe der einzelnen Texte) der Parameter a gegen 0 – was insofern von Interesse ist, als sich für $a = 0$ die (1-verschobene) geometrische Verteilung als ein Spezialfall der Good-Verteilung erweist (Wimmer/Altmann 1999: 219f.).

Tab. 4a/b: Anpassung der rechts-gestutzten Zeta-Verteilung und der Zipf-Mandelbrot-Verteilung an 30 slowakische Texte

	rechts-gestutzt Zeta, *R*=46			**Zipf-Mandelbrot (*a,b*) *n*=46**			
Nr.	**a**	$\chi^2_{FG\ 39}$	**C**	**a**	**b**	$\chi^2_{FG=28}$	**C**
1	0,8030	1484,35	0,2139	11,42	98,96	258,78	0,0373
2	0,7987	3611,14	0,2182	12,00	106,51	665,78	0,0402
3	0,8038	920,04	0,2240	5,92	43,87	174,16	0,0424
4	0,8043	727,91	0,2158	9,85	81,53	140,72	0,0417
5	0,8024	1912,52	0,2258	10,03	82,21	323,77	0,0382
6	0,8350	1471,38	0,2094	2,95	15,03	355,62	0,0506
7	0,8120	4754,68	0,2007	5,11	36,15	1034,45	0,0437
8	0,7992	2115,74	0,2036	8,03	66,43	511,68	0,0492
9	0,8150	2706,31	0,1980	2,02	7,95	965,54	0,0706
10	0,8267	1145,97	0,2161	12,00	111,70	160,61	0,0303
11	0,7660	528,23	0,2128	6,34	51,90	98,81	0,0398
12	0,7770	1434,92	0,2414	7,51	61,18	255,28	0,043
13	0,7730	1236,79	0,2211	7,18	59,20	222,01	0,0397
14	0,8028	845,02	0,2223	12,00	104,24	114,75	0,0302
15	0,7989	1172,42	0,2274	1,82	6,29	489,26	0,0949
16	0,8008	396,11	0,2466	5,37	37,76	78,48	0,0489
17	0,8063	748,12	0,2127	1,56	4,43	352,05	0,1001
18	0,7771	378,34	0,2574	1,49	4,11	202,21	0,1376

19	0,7949	413,97	0,2576	6,40	46,15	68,91	0,0429
20	0,7924	501,19	0,2037	12,00	119,68	80,64	0,0328
21	0,9011	128,78	0,2292	5,49	32,52	33,61	0,0598
22	0,8209	104,65	0,2362	12,00	118,13	32,89	0,0742
23	0,8290	130,16	0,2059	1,35	2,68	78,49	0,1242
24	0,8071	212,31	0,1749	1,99	8,20	95,75	0,0789
25	0,8246	284,10	0,2123	12,00	110,81	72,38	0,0541
26	0,8014	313,58	0,1847	1,68	5,45	130,18	0,0767
27	0,7576	347,69	0,2346	12,00	118,94	64,75	0,0437
28	0,7749	964,16	0,2298	11,99	109,94	139,54	0,0333
29	0,7607	309,14	0,2389	1,91	7,97	129,60	0,1002
30	0,8066	363,39	0,2633	5,47	36,99	65,90	0,0478

Tab. 5a/b: Anpassung der rechts-gestutzten geometrischen Verteilung und der Good-Verteilung an 30 slowakische Texte

	rechts-gestutzt geometrisch, (*q*, *R* = 46)			**Good-1 (*a*, *p*)**			
Nr.	**q**	$\chi^2_{FG=43}$	**C**	**a**	**p**	$\chi^2_{FG=43}$	**C**
1	0,9063	232,90	0,0336	0,000007	0,9039	230,20	0,0332
2	0,905	587,59	0,0355	0,752803	0,9800	2709,98	0,1638
3	0,9051	115,89	0,0282	0,000007	0,9028	114,31	0,0278
4	0,9047	116,58	0,0346	0,000012	0,9024	338,11	0,0247
5	0,904	253,81	0,0300	0,000000	0,9017	250,58	0,0296
6	0,9004	147,45	0,0210	0,784600	0,9800	1076,63	0,1532
7	0,9045	722,00	0,0305	0,000027	0,9023	712,20	0,0301
8	0,9072	446,20	0,0429	0,000012	0,9047	442,46	0,0426
9	0,9047	343,30	0,0251	0,000012	0,9024	338,11	0,0247
10	0,9022	123,10	0,0232	0,000001	0,9001	120,98	0,0228
11	0,9114	70,97	0,0286	0,723979	0,9800	394,00	0,1587
12	0,9084	179,05	0,0301	0,000000	0,9057	177,09	0,0298
13	0,9098	154,96	0,0277	0,000002	0,9070	153,35	0,0274
14	0,9049	97,55	0,0257	0,757333	0,9800	625,64	0,1646
15	0,9039	111,11	0,0216	0,000001	0,9017	109,12	0,0212
16	0,9039	46,11	0,0287	0,000130	0,9015	45,61	0,0284
17	0,9051	82,84	0,0236	0,759283	0,9800	551,76	0,1569
18	0,9068	46,46	0,0316	0,000002	0,9041	45,97	0,0313
19	0,9025	35,58	0,0221	0,000011	0,9002	34,99	0,0218
20	0,9076	61,68	0,0251	0,000000	0,9050	60,78	0,0247
21	0,8866	26,40	0,0470	0,844402	0,9800	97,01	0,1726
22	0,9039	27,94	0,0631	0,769155	0,9800	81,50	0,1840
23	0,9001	22,29	0,0353	0,030432	0,9009	21,51	0,0340

24	0,9068	58,19	0,0479	0,755326	0,9800	158,68	0,1307
25	0,8991	56,58	0,0423	0,775746	0,9800	213,71	0,1597
26	0,9065	30,50	0,0180	0,754023	0,9800	225,28	0,1327
27	0,9104	53,84	0,0363	0,717663	0,9800	264,74	0,1786
28	0,9108	121,33	0,0289	0,731983	0,9800	721,85	0,1720
29	0,9107	42,85	0,0331	0,720814	0,9800	234,25	0,1810
30	0,9013	34,22	0,0248	0,762362	0,9800	272,48	0,1974

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 slowakischer 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. Auf rangierte Graphemhäufigkeiten ist sie bislang nur vereinzelt angewendet worden. Erstmals hat sie Grzybek (2001) auf der Basis eines Textes von A.S. Puškin („Царь Салтан“) ins Spiel gebracht; das mit einem Wert von $C = 0.0082$ ausgezeichnete Anpassungsergebnis konnten. Eine erste systematische Untersuchung zur Eignung der negativ hypergeometrischen Verteilung für Buchstabenhäufigkeiten war die oben erwähnte Studie von Grzybek/Kelih/Altmann (2004), in der die negativ hypergeometrische Verteilung auf der Basis einer größeren Anzahl russischer Texte im Vergleich zu anderen Modellen getestet und als überaus geeignetes Modell nachgewiesen wurde. Bei dieser systematischen Untersuchung nicht nur von individuellen Texten, sondern auch von Textsegmenten, Textkumulationen, und Textmischungen konnte das Einzelergebnis von Grzybek (2001) für das Russische auch auf breiterer Basis bestätigt werden. In ähnlicher Weise wies Best (2003: 79) zunächst am Beispiel einer kurzen Fabel von Pestalozzi auf die Eignung der negativ hypergeometrischen Verteilung für die Modellierung der Ranghäufigkeit deutscher Buchstaben hin, bevor er diese Beobachtung an einer größeren Anzahl deutscher Texte bestätigen konnte Best (2005). Dabei

hat Best – um die Eignung der negativ hypergeometrischen Verteilung zu unterstreichen – in den einzelnen Datensätzen Leer-, Formatierungs- und Sonderzeichen ebenso wie diakritische Zeichen unterschiedlich gehandhabt, wodurch freilich ein interpretierender Vergleich der Daten zueinander und mit anderen Sprachen nur in eingeschränktem Maße möglich ist.
Tab. 6 veranschaulicht die Güte der Anpassung, wie sie sich für das Gesamtkorpus der slowakischen Texte ergibt: die sich aufgrund der angeführten Parameter ergebenden theoretischen Häufigkeiten NP_i sind neben den beobachteten Häufigkeiten f_i der 46 slowakischen rangierten Grapheme dargestellt; Abb. 3 repräsentiert die Ergebnisse in anschaulicher Form. Wie den Daten der Tab. 6 sowie der Abb. 3 zu entnehmen ist, stellt sich in der Tat die negativ hypergeometrische Verteilung als ein gutes Modell dar (C = 0.0139).

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

i	***f*(i)**	***NP*(i)**	***i***	***f*(i)**	***NP*(i)**
1	14194	15621,77	**24**	1611	1673,05
2	13772	12353,98	**25**	1593	1502,40
3	12701	10746,13	**26**	1465	1343,42
4	9285	9622,56	**27**	1422	1195,61
5	8323	8733,81	**28**	1395	1058,51
6	7099	7986,54	**29**	1294	931,70
7	6562	7335,76	**30**	1073	814,76
8	6534	6756,26	**31**	947	707,31
9	6164	6232,47	**32**	719	608,99
10	6091	5754,03	**33**	402	519,45
11	5731	5313,71	**34**	346	438,33
12	5659	4906,21	**35**	297	365,32
13	5103	4527,52	**36**	270	300,08
14	4121	4174,53	**37**	253	242,29
15	3845	3844,75	**38**	172	191,64
16	3135	3536,17	**39**	131	147,80
17	2676	3247,10	**40**	124	110,44
18	2660	2976,13	**41**	47	79,24
19	2408	2722,07	**42**	27	53,85
20	2262	2483,88	**43**	10	33,91
21	1954	2260,64	**44**	3	19,02

22	1825	2051,56	**45**	2	8,76
23	1685	1855,91	**46**	0	2,62
	K = 4,1588			χ^2 = 2053,13	
	M = 0,8317			FG = 42	
	n = 45			C = 0,0139	

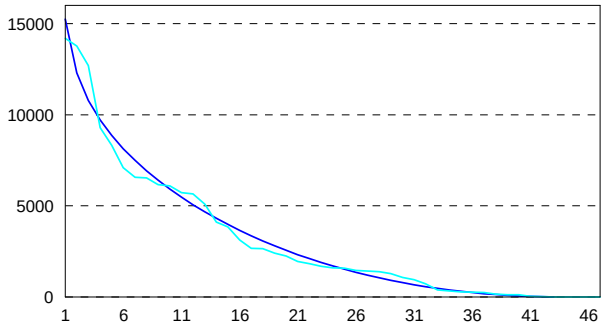

Abb. 3:
Anpassung der neg. hypergeometrischen Verteilung an 30 slowakische Texte

Tab. 7 zeigt die Ergebnisse der Anpassungen für alle einzelnen Stichproben; es bestätigen sich auch hier die guten Ergebnisse: Der Diskrepanzkoeffizient liegt insgesamt im Intervall von 0.0423; ≥ C ≥ 0.008; dabei beläuft er sich in 25 der 30 Einzelanalysen auf einen Wert von $C < 0.02$, davon wiederum beträgt er in 5 der Fälle $C < 0.01$. Abb. 4 veranschaulicht die Werte des Diskrepanzkoeffizienten C für die 30 slowakischen Texte.

Tab. 7: Ergebnis der Anpassung der negativ hypergeometrischen Verteilung an 30 slowakische Texte

Neg. Hypergeometrisch				
Nr.	**K**	**M**	$\chi^2_{FG=42}$	**C**
1	4,1297	0,8178	146,45	0,0211
2	4,2204	0,8292	258,53	0,0156
3	4,3503	0,8534	57,02	0,0139

4	4,2585	0,8348	52,82	0,0157
5	4,3852	0,8515	102,71	0,0121
6	4,3476	0,8208	69,50	0,0099
7	4,1721	0,8157	272,77	0,0115
8	4,0558	0,8086	220,91	0,0213
9	4,1453	0,8130	144,88	0,0106
10	4,3351	0,8295	73,51	0,0139
11	4,0793	0,8483	29,93	0,0121
12	4,5016	0,9052	74,15	0,0125
13	4,2927	0,8782	53,35	0,0095
14	4,3227	0,8447	45,13	0,0119
15	4,2485	0,8335	4,25	0,0008
16	4,6526	0,8979	20,01	0,0125
17	4,2608	0,8360	37,79	0,0107
18	4,5578	0,9048	22,55	0,0153
19	4,8046	0,9226	10,58	0,0066
20	4,1325	0,8329	24,75	0,0101
21	4,8618	0,7886	12,81	0,0228
22	4,3766	0,8268	18,72	0,0423
23	4,2510	0,7976	7,58	0,0120
24	3,8135	0,7554	16,40	0,0135
25	4,3137	0,7993	15,47	0,0116
26	3,9301	0,7920	8,48	0,0050
27	4,3933	0,8995	16,74	0,0113
28	4,2089	0,8686	87,73	0,0209
29	4,3783	0,8996	16,92	0,0131
30	4,6310	0,8847	22,95	0,0166

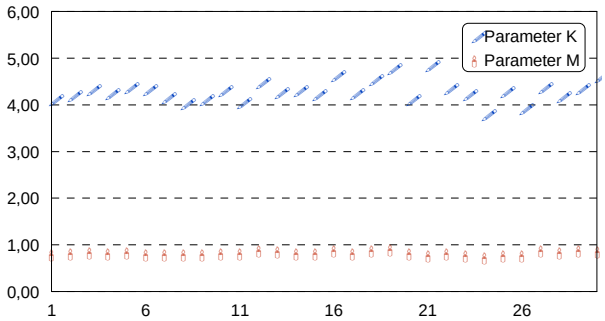

Abb. 4:
Diskrepanzkoeffizient *C* (neg. hypergeom. Verteilung)

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 relativ stabil; auch die Parameter *K* und *M* stellen sich als ziemlich konstant dar: Abgesehen von dem ohnehin konstanten Parameter *n* (der konstant bei $n = 45$, d.h. um eins niedriger als der Inventarumfang liegt), liegen die Werte für *K* im Intervall von $4.86 \geq K \geq 3.81$, wobei das 95%-Konfidenzintervall bei einer Unter- bzw. Obergrenze von $K_u = 4.23$ und $K_o = 4.40$ relativ eng ist. Dasselbe gilt für den Parameter *M*, der im Intervall zwischen und $0.92 \geq M \geq 0.76$ liegt, wobei auch hier das 95%-Konfidenzintervall bei einer Unter- bzw. Obergrenze von $M_u = 0.83$ und $M_o = 0.86$ sehr schmal ist. Abb. 5 veranschaulicht die relative Konstanz der Ergebnisse.

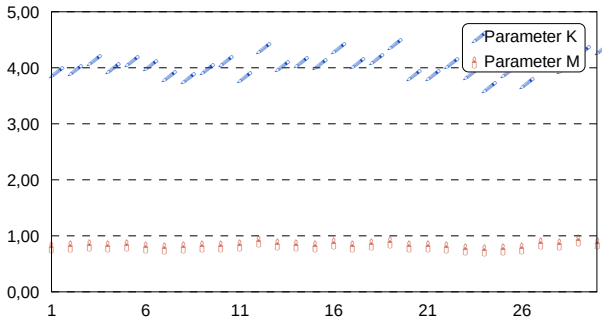

Abb. 5:
Konstanz der Parameter *K* und *M*

Eine genauere Inspektion der Abb. 5 legt allerdings einen interessanten Befund nahe: Es schaut nämlich fast so aus, als ob Parameter *K* und *M* nicht wirklich – wie dies auch in den Analysen von Grzybek/Kelih/Altmann (2004, 2005b) geschlussfolgert wurde – konstant sind; eher entsteht der Eindruck, als ob Schwankungen des einen Parameters (die insgesamt freilich relativ gering sind) auch Schwankungen des anderen Parameters nach sich ziehen. In der Tat bestätigt die Berechnung des Korrelationskoeffizienten (der aufgrund der gegebenen Normalverteilung sowohl von *K* als auch von *M* zulässig ist), einen solchen Zusammenhang, der bei $r = 0.592$ hoch signifikant ist ($p = 0.001$). Dieser erstmalig beobachtete Befund ist insofern von Interesse, als ein allfälliger Zusammenhang zwischen den beiden Parametern *K* und *M* erste Hinweise auf eine mögliche Interpretation bietet, der es in den

abschließenden Zusammenfassung noch einmal ausführlicher nachzugehen gilt.

In Anbetracht der Tatsache, dass die negativ hypergeometrische Verteilung drei Parameter aufweist, von denen einer (n) direkt vom Inventarumfang abhängt, ist zuvor von Interesse, ob sich auch die Whitworth-Verteilung für slowakische Graphemhäufigkeiten als ein geeignetes Modell erweist. Dies ist zum einen deshalb von besonderem Interesse, weil die Whitworth-Verteilung ja nur einen Parameter hat, der zudem ausschließlich durch den Inventarumfang vorgegeben ist, zum anderen, weil sie sich im Falle der russischen Grapheme als überaus zufrieden stellendes Modell erweist (vgl. Grzybek/Kelih/Altmann 2004). Tab. 8 stellt die Ergebnisse zu den slowakischen Texten im Detail dar. Deutlich ist zu sehen, dass die Whitworth-Verteilung sich nicht für die Modellierung der slowakischen Daten eignet: Der Diskrepanzkoeffizient liegt für die einzelnen Stichproben im Intervall $0.1608 \geq C \geq 0.0516$.

Tab. 8: Ergebnis der Anpassung der Whitworth-Verteilung an 30 slowakische Texte

W h i t w o r t h, R = 46					
Nr.	$\chi^2_{FG=44}$	**C**	**Nr.**	$\chi^2_{FG=44}$	**C**
1	509,06	0,0734	**16**	137,79	0,0858
2	1247,25	0,0754	**17**	244,36	0,0695
3	308,14	0,0750	**18**	118,86	0,0809
4	258,56	0,0767	**19**	139,04	0,0865
5	662,15	0,0782	**20**	146,60	0,0596
6	587,16	0,0836	**21**	90,39	0,1608
7	1678,02	0,0708	**22**	47,75	0,1078
8	741,51	0,0714	**23**	55,24	0,0874
9	927,48	0,0679	**24**	77,14	0,0635
10	433,79	0,0818	**25**	124,39	0,0930
11	132,25	0,0533	**26**	87,62	0,0516
12	427,98	0,0720	**27**	95,58	0,0645
13	333,99	0,0597	**28**	271,59	0,0647
14	278,53	0,0733	**29**	84,74	0,0655
15	375,39	0,0728	**30**	132,10	0,0957

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. 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 für die Modellierung der Vorkommenshäufigkeit slowakischer Grapheme nicht eignen; dieser Befund deckt sich mit den Ergebnissen zum Russischen (Grzybek/Kelih/Altmann 2004) sowie zum Slowakischen unter Annahme eines Inventars von 43 Graphemen (Grzybek/Kelih/Altmann 2005b). Es liegt deshalb nahe, dass in dieser Hinsicht eine Reihe von Annahmen auch im Hinblick auf andere Sprachen zu korrigieren sein werden – das jedoch bedarf weiterer empirischer Überprüfungen.
2. Die Whitworth-Verteilung führt – ebenso wie im Fall der Annahme eines 43 Grapheme umfassenden Bestands – im Fall des Slowakischen zu keinen befriedigenden Resultaten; dies ist ein klarer Gegensatz im Vergleich zu den Befunden zum Russischen.
3. Als ein geeignetes Modell für die Modellierung rangierter Graphemhäufigkeiten des Slowakischen eignet sich die negativ hypergeometrische Verteilung, mit der sich gute Anpassungsergebnisse erzielen lassen. Allerdings hat diese Verteilung nicht weniger als drei Parameter (n, K, M), von denen sich nur einer (n) aufgrund seiner Abhängigkeit vom Inventarumfang direkt interpretieren lässt.
4. Unter vergleichendem Einbezug der Ergebnisse zum Russischen und Slowakischen (mit $n = 43$) bieten die Ergebnisse der vorliegenden Studie allerdings erstmals Hinweise im Hinblick auf eine Interpretation auch der Parameter K und M. Denn aufgrund der bisherigen Befunde war davon auszugehen, dass diese innerhalb der Texte einer gegebenen Sprache eine relative Konstanz aufweisen. Allerdings ergab bereits der Vergleich der Ergebnisse zum Russischen und Slowakischen (mit $n = 43$) dass der Parameter K für das Slowakische deutlich höher war als für das Russische, während der Parameter M sich nicht wesentlich von dem im Russischen unterscheidet. Diesem Umstand – der von Grzybek/Kelih/Altmann (2005b) als ein Indiz dafür in Betracht gezogen wurde, dass die Parameter K und M in zumindest indirekter Abhängigkeit vom Inventarumfang zu interpretieren sein könnten – wird es näher nachzugehen sein.

a. Ein in Form des *t*-Tests durchgeführter Mittelwertvergleich der Parameterwerte des Slowakischen unter Berücksichtigung der unterschiedlichen Inventargröße ($n = 43$ vs. $n = 46$) zeigt, dass der Parameter *K* für $n = 46$ signifikant größer ist als für $n = 43$ ($t_{FG=56} = 4{,}53$; $p < 0{,}001$). Damit bestätigt sich die schon im Vergleich zum Russischen beobachtete Tendenz, die allerdings eine detailliertere Re-Analyse des Russischen erfordert, und zwar (a) unter Berücksichtigung ausschließlich homogener Texte und (b) der Tatsache, dass auch im Russischen – je nach Berücksichtigung des Graphems ë – die Inventargröße per definitionem zwischen $n = 32$ und $n = 33$ variieren kann.
b. Ein in Form des *t*-Tests durchgeführter Mittelwertvergleich der Parameterwerte des Slowakischen unter Berücksichtigung der unterschiedlichen Inventargröße ($n = 43$ vs. $n = 46$) zeigt, dass der Parameter *K* sich für beide Bedingungen nicht signifikant unterscheidet ($t_{FG=58} = 1{,}07$; $p = 0{,}29$).
c. Die Analyse des Zusammenhangs zwischen den beiden Parametern *K* und *M* hat in der vorliegenden Studie des Slowakischen (also unter der Annahme von $n = 46$) ergeben, dass eine hochgradig signifikante Korrelation zwischen K und M besteht ($r = 0.59$, $p = 0{,}001$); diese Tendenz wird durch eine Re-Analyse des Slowakischen unter der Annahme von $n = 43$ bekräftigt, wo der Zusammenhang noch stärker ausgeprägt ist ($r = 0.83$, $p < 0{,}001$).

Damit ergibt sich eine vielversprechende Perspektive, insofern eine Interpretation beider Parameter (*K* und *M*) in greifbare Nähe rückt: Diese Interpretation könnte *K* als abhängig vom der jeweiligen Inventarumfang einer Sprache, und *M* innerhalb einer gegebenen Sprache als abhängig von *M* beinhalten. Eine Überprüfung dieser Annahme kann jedoch nur in weiteren empirischen Studien an Sprachen mit unterschiedlichem Umfang des Grapheminventars vorgenommen werden. Entsprechende Untersuchungen für weitere slawische Sprachen sind bereits in Arbeit: Von besonderem Interesse wird es dabei sein, vergleichende Untersuchungen zum Slowenischen durchzuführen, das ja mit $n = 25$ Graphemen das Minimum unter den slawischen Sprachen repräsentiert (vgl. Grzybek/Kelih/Altmann 2005a), wohingegen die vorliegende Studie zum Slowakischen mit $n = 46$ das Maximum darstellt.

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 und Slowakische hinausgehender Relevanz sind. Nicht zuletzt ergeben sich so Einsichten in die graphematischen Strukturen verschiedener (slawischer) Sprachen, incl. historisch-diachronischer Fragen orthographischer Natur.

5. 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.[4]
6. Wie die Analyse der slowakischen Texte zeigt, scheinen schlechte Anpassungsergebnisse insbesondere bei kurzen Texten vorzukommen; dies macht es erforderlich, dem Faktor der (notwendigen bzw. optimalen) Stichprobengröße in Zukunft systematisch nachzugehen (vgl. Grzybek/Kelih/Altmann 2005c).
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 2005c).

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.

---

[4] 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.

Best, K.H. (2003): *Quantitative Linguistik: Eine Annäherung*. 2., überarbeitete und erweiterte Auflage. Göttingen.

Best, K.H. (2005): Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten [In Vorb.]

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

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 slawische Philologie, XXXI;* 131-162.

Grzybek, P.; Kelih, E.; Altmann, G. (2004): „Graphemhäufigkeiten (am Beispiel des Russischen. Teil II: Modelle der Häufigkeitsverteilung", in: *Anzeiger für slawische Philologie, XXXII;* 25-54.

Grzybek, P.; Kelih, E.; Altmann, G. (2005a): „Graphemhäufigkeiten im Slowenischen." [In print]

Grzybek, P.; Kelih, E.; Altmann, G. (2005b): „Graphemhäufigkeiten im Slowakischen. (Teil I: Ohne Digraphen)". In: Nemcová, E. (Hrsg.), *Philologia actualis slovaca*. [Im Druck].

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

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: de Gruyter, 514-546.

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.

Wimmer, G.; Altmann, G. (1999): *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

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. (2005a): „Unified Derivation of Some Linguistic Laws." In: *Handbook of Quantitative Linguistics*. [In print]

Wimmer, G.; Altmann, G. (2005b): „Towards a Unified Derivation of Some Linguistic Laws." In: Grzybek, P. (ed.), *Word Length Studies and Related Issues*. [In print]

Wimmer, G.; Wimmerová, S. (Ms.): „Ein musikalisches Rangordnungsgesetz."

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.