Проблеми квантитативної лінгвістики, 2005, с. 159-179.
Problems of Quantitative Linguistics, 2005, pp. 159-179.

Graphemhäufigkeiten im Ukrainischen Teil I: Ohne Apostroph (')

Peter Grzybek (Graz, Österreich), Emmerich Kelih (Graz, Österreich)*

METHODOLOGISCHE VORBEMERKUNGEN

In der vorliegenden Untersuchung geht es um die Vorkommenshäufigkeit von Graphemen im Ukrainischen. Es handelt sich hierbei um den Teil einer Serie von Untersuchungen zur Vorkommenshäufigkeit von Graphemen in verschiedenen (slawischen) Sprachen. Den Auftakt zu dieser Untersuchungsserie stellte eine historische Darstellung zur Erforschung von Graphemhäufigkeiten des Russischen dar (Grzybek/Kelih 2003), begleitet von einer Reihe allgemeiner methodologischer Bemerkungen. Als eines der Ergebnisse stellte sich bei der synoptischen Darstellung deutlich heraus, dass es bei allen Untersuchungen eigentlich nie um die einfache Erhebung von Buchstabenhäufigkeiten an und für sich ging; vielmehr waren mit allen Studien immer auch weiterführende Fragen verbunden, 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.

Dabei hat es allerdings auch immer ein übergeordnetes Interesse an der Vergleichbarkeit unterschiedlicher Datenerhebungen gegeben. Diese Frage nach der Vergleichbarkeit war entweder ausgerichtet auf die „Repräsentativität" einer Stichprobe und damit primär auf die Frage der (minimal notwendigen) Größe einer Stichprobe, oder aber auf die Bestim-

* Address correspondence to: Peter Grzybek. Universität Graz, Institut für Slawistik, A-8010 Graz. Email: peter.grzybek@uni-graz.at.

Diese Publikation steht im Zusammenhang mit Grazer Datenbank, die im Rahmen des vom österreichischen Fonds für wissenschaftliche Forschung (FWF) geförderten Projekts P15485 („Wortlängenhäufigkeiten in slawischen Texten") erstellt wurde; vgl.: http://www-gewi.uni-graz.at/quanta. Sie entstand außerdem im Rahmen eines DOC-Stipendium der Österreichischen Akademie der Wissenschaften für Emmerich Kelih.

mung von spezifischen und generellen Charakteristiken der untersuchten Daten. Immer wieder ging es vor allem um den Vergleich der Vorkommenshäufigkeit einzelner Grapheme oder bestimmter Graphemgruppen (wie z.B. vokalischer vs. konsonantischer, o.ä.) unter besonderer Berücksichtigung vermeintlich unterschiedlichen Datenmaterials, etwa der Gegenüberstellung von mündlichen vs. schriftlichen Quellen, poetischer vs. prosaischer Texte, wissenschaftlicher vs. journalistischer Texte.

Der Frage eines einheitlichen Häufigkeitsmodells, welches verschiedenen Stichproben ungeachtet der Häufigkeit der individuellen Grapheme zugrunde liegt, ist in der Vergangenheit sehr viel seltener gestellt worden. Ohne im Detail auf die methodologischen Aspekte einzugehen, sei zusammenfassend herausgestellt, dass sich das Interesse dieser Art von Untersuchung nicht auf die Häufigkeit der jeweils einzelnen Grapheme ausrichtet; vielmehr wird die Frage gestellt, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, zum dritthäufigsten, usw. hat. Untersucht werden also so genannte Rang-Häufigkeitsverteilungen – damit ist 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 aufwiesen oder nicht.

Einhergehend mit einer methodologischen Diskussion dieser Arbeiten wurde diese Problematik im Anschluss an die erwähnte Arbeit von Grzybek/Kelih (2003) dann von Grzybek, Kelih, und Altmann zunächst am Russischen (2004), später dann auch am Slowakischen[*] (2005a,b) und Slowenischen (2005c) untersucht. An diese Untersuchungen knüpft

[*] Zum Slowakischen wurden zwei Studien durchgeführt, und zwar in Abhängigkeit von der Behandlung der Digraphen DZ, DŽ, und CH: In der ersten der beiden (2005a) wurden keine Digraphen zugrundegelegt, so dass der Inventarumfang $n = 43$ betrug, in der zweiten wurden die Digraphen als eigene Klasse gewertet, was entsprechend in einem Inventarumfang von $n = 46$ resultiert.

die vorliegende Studie an: Es sollen erstmals entsprechende Untersuchungen zum Ukrainischen vorgestellt werden. Dabei soll die Suche nach einem einheitlichen, verschiedenen Stichproben gemeinsam zugrunde liegenden formalisierten Modell im Vordergrund stehen.

Diese Behandlung dieser Frage wird den Verlauf der folgenden Überlegungen prägen, wobei wir unsere Analysen an einem für diese Zwecke zusammengestellten Korpus von 30 ukrainischen Texte durchführen werden (s.u.). Dabei legen wir den folgenden Analysen einen Alphabet- bzw. Inventarumfang von 33 Graphemen zugrunde:

```
а,  б,  â,  ã,  ´,  ä,  å,  °,  æ,  ç,  è,  ³,
                ¿,  é,  ê,  ë,  ì,
í,  î,  ï,  ð,  ñ,  ò,  ó,  ô,  õ,  ö,  ÷,  ø,
                ù,  þ,  ÿ,  ü
```

Keine Berücksichtigung findet somit der Apostroph ('), der in der schriftsprachlichen Praxis in der Regel zwischen Konsonanten und Vokalen verwendet wird zur Markierung der nicht-palatalisierten Aussprache des betreffenden Konsonanten; da dieser allerdings überwiegend als nicht zum Alphabetsystem gehörig betrachtet wird, bleibt er von der vorliegenden Untersuchung ausgeschlossen. Ungeachtet dessen wird es an anderer Stelle notwendig sein, die hier folgenden Analysen im Hinblick auf die offensichtliche Relevanz des Apostroph auch unter dessen Berücksichtigung als Systemelement zu reproduzieren.

Wenden wir uns damit den skizzierten Analysen zu, wobei eingangs eine Behandlung der theoretischen Aspekte von in den einschlägigen Diskussionen bislang ins Spiel gebrachten Modellen zu leisten ist.

MODELLE ZUR ERFASSUNG VON GRAPHEMHÄUFIGKEITEN

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*, 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
(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.

EMPIRISCHE ÜBERPRÜFUNG DER MODELLE

Text- und Datenbasis

Bei der empirischen Überprüfung der oben dargestellten Modelle an ukrainischen 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 ukrainischem) Datenmaterial getestet werden. Während allerdings in der genannten Untersuchung zum Russischen von Grzybek/Kelih/Altmann (2004) zu diesem Zweck systematisch und kon-

* 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.

trolliert 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.

Tabelle 1

Text- und Datenbasis

No.	Autor	Titel	Textsorte	N*
1	Artem Sacharčenko	Postfol'k	Drama	5682
2	Anastasija Rižakova	Sumne vesilja		3053
3	Roman Kucharuk	Portret (pesa)		24318
4	Anna Bagrjana	Nad časom		7064
5	anonym	Ja ne bažaju pani Timošenko [...]	Publizistik	6372
6	anonym	Evropa – Ukrajina		5682
7	anonym	Novi členi urjadu [...]		7452
8	anonym	Nas svonu učikue [...]		6912
9	anonym	Vijna i Ukrajina		6362
10	Anna Chromova	Tapeti	Lyrik	6345
11	Jurij Andrjuchovič	Listi v Ukrajinu		9051
12	Volodimir Ljaškevič	Palac		6867
13	Viktorija Lemeševa	Mij ljubij eve	Prosa	9406
14	Dimitro Bondarenko	Čerešni		6724
15	Volodimir Javoriv'skij	Vovča Ferma		17030
16	Ostap Sokoljuk	Kolip očej tvojich		15708
17	Jurij Šeljaženko	Razom!		17935
18	Volodimir Viničenko	Božki 1		38288
19	Volodimir Viničenko	Božki 2		7527
20	Volodimir Viničenko	Božki 3		22101
21	Viktor Petrov	Doktor Serafiks 1		14983
22	Viktor Petrov	Doktor Serafiks 2		13575
23	Viktor Petrov	Doktor Serafiks 3		16945
24	Ol'ga Kobijans'ka	V nedeliju rano 1 [...]		24246
25	Ol'ga Kobijans'ka	V nedeliju rano 2 [...]		27801
26	Ol'ga Kobijans'ka	V nedeliju rano 3 [...]		26085

* N ist hier die absolute Häufigkeit der Grapheme in den jeweiligen Texten.

27	Mikola Zerov	Nove Ukrajins'ke pis'menstvo 1	Wissenschaft	7877
28	Mikola Zerov	Nove Ukrajins'ke pis'menstvo 2		8242
29	Mikola Zerov	Nove Ukrajins'ke pis'menstvo 3		5488
30	Sergej Evremov	Istoriji ukrajins'kogo pis'menstva		11495
	Gesamt			**386616**

In der Tabelle 1 findet sich im Anschluss an die jeweilige 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, eine Zuordnung zu einem Texttyp[*], und schließlich der Umfang des Textes (in der Anzahl der Buchstabenvorkommnisse). In der letzten Zeile der Tabelle finden sich die aufsummierten Werte eines aus den 30 einzelnen Texten zusammengestellten Textkorpus.

Wie oben bereits gesagt wurde, soll es in der vorliegenden Untersuchung primär um die individuellen Texte, wie sie in Tab. 1 aufgeschlüsselt sind, nicht um das gesamte Korpus gehen. Dennoch lässt sich das Vorgehen exemplarisch am Gesamtkorpus der Texte veranschaulichen: Fügt man die 30 einzelnen Texte zu einem Gesamtkorpus zusammen, so beläuft sich der Umfang dieses Korpus auf $N = 386616$ Grapheme. Tab. 2 gibt die absoluten und relativen Häufigkeiten für die 33 Grapheme des Ukrainischen wieder.

Tabelle 2

Graphemhäufigkeiten im Gesamtkorpus

Graphem	f_i	f_i(%)	Graphem	f_i	f_i(%)
а	32774	8,4771	і	24639	6,3730
б	7487	1,9365	î	37267	9,6393
â	21053	5,4455	ï	10584	2,7376
ã	6406	1,6569	ð	16240	4,2006
´	78	0,0202	ñ	16296	4,2150
ä	12949	3,3493	ò	20075	5,1925
å	19171	4,9587	ó	12959	3,3519
°	2407	0,6226	ô	506	0,1309
æ	3876	1,0025	õ	4625	1,1963

[*] Auch zu der Frage der Text-Typologisierung kann an dieser Stellen nicht detailliert Stellung bezogen werden – vgl. hierzu: Gryzbek/Kelih 2005, Grzybek/Stadlober/ Kelih/Antić (2005), Kelih/Antić/Grzybek/Stadlober 2005, Grzybek/Kelih/Stadlober (2005).

ç	8944	2,3134	ö	2857	0,7390
è	25080	6,4871	÷	5850	1,5131
³	20941	5,4165	ø	3565	0,9221
¿	2790	0,7216	ù	2484	0,6425
é	5074	1,3124	þ	3843	0,9940
ê	13697	3,5428	ÿ	8877	2,2961
ë	13936	3,6046	ü	6888	1,7816
ì	12398	3,2068	**Gesamt**	**386616**	**1**

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

Um die theoretische Modellierung der Ranghäufigkeiten in den 30 ukrainischen Texten geht es in den folgenden Analysen.*

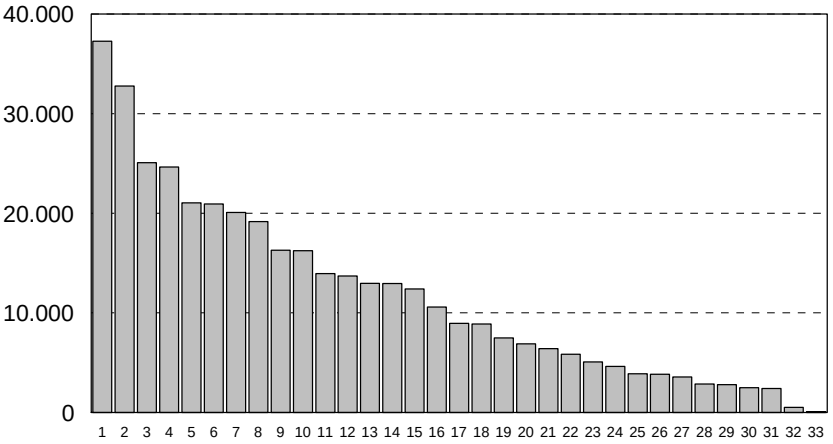

Abb. 1
Ranghäufigkeitsverteilung ukrainischer Grapheme (Gesamtkorpus)

ERGEBNISSE

Schauen wir uns im folgenden die Ergebnisse für die oben diskutierten Verteilungsmodelle im einzelnen an. Die Grundidee besteht darin, die

* 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.

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 rechts-gestutzte Zeta-Verteilung an die Daten des Gesamtkorpus an. Tab. 3 enthält neben den Rängen 1 bis 33 die absoluten Häufigkeiten *f(i)* in absteigender Reihenfolge. Der Parameter $R = 33$ 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 $a = 0{,}6375$. Setzt man diese Werte in die Formel (1) ein, ergeben sich die theoretischen Werte *NP(i)*.

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

Wie den Werten in Tab. 3 und ihrer Veranschaulichung in Abb. 2 zu entnehmen ist, stellt die Zeta-Verteilung für die Daten des Gesamtkorpus kein gutes Modell dar; dies drückt sich auch im Wert des Diskrepanzkoeffizienten $C = 0{,}1064$ klar aus. 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 (4a/b) enthält Tab. 4a/b auch die Anpassungsergebnisse der Zipf Mandelbrot-Verteilung.

Tabelle 3

Ergebnis der Anpassung der rechts-gestutzten Zeta-Verteilung an das Gesamtkorpus

i	f_i	***NP*(i)**	**i**	f_i	***NP*(i)**
1	37267	50585,06	**18**	8877	8011,76
2	32774	32516,34	**19**	7487	7740,30
3	25080	25109,34	**20**	6888	7491,27
4	24639	20901,68	**21**	6406	7261,83
5	21053	18129,95	**22**	5850	7049,62
6	20941	16140,42	**23**	5074	6852,64
7	20075	14629,62	**24**	4625	6669,20
8	19171	13435,71	**25**	3876	6497,87
9	16296	12463,74	**26**	3843	6337,40
10	16240	11654,02	**27**	3565	6186,74
11	13936	10966,96	**28**	2857	6044,94
12	13697	10375,15	**29**	2790	5911,20
13	12959	9858,97	**30**	2484	5784,81
14	12949	9404,00	**31**	2407	5665,13
15	12398	8999,32	**32**	506	5551,62
16	10584	8636,54	**33**	78	5443,76
17	8944	8309,10			
$a = 0{,}6375$				$\chi^2 = 41138{,}90$	
$R = 33$				FG = 30	
$N = 386616$				C = 0,1064	

Deutlich ist zu sehen, dass beide Verteilungen, die in der Vergangenheit wiederholt in Betracht gezogen worden sind, sich nicht für die Modellierung der Ranghäufigkeit ukrainischer Grapheme eignen: Bei der Zeta-Verteilung liegen die Werte des Diskrepanzkoeffizienten für die einzelnen Datensätze im Intervall von $0.1409 \geq C \geq 0.0879$, für das gesamte Korpus beträgt der Wert $C = 0.1064$ 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 (ukrainische) Graphemhäufigkeiten eignet: Von den 30 Stichproben kommen nur zwei Texte auf einen Wert von $C < 0.02$, wobei der Diskrepanzkoeffizienten für das Gesamtkorpus $C = 0.0602$ beträgt.

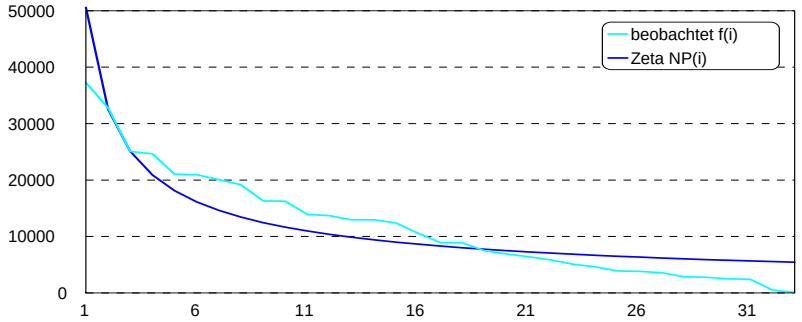

Abb. 2
Anpassung der rechts-gestutzten Zeta-Verteilung (Gesamtkorpus)

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 ukrainischer Graphemhäufigkeiten: Die Werte des Diskrepanzkoeffizienten C liegen für die einzelnen Stichproben bei der geometrischen Verteilung im Intervall von $0.0314 \geq C \geq 0.0149$, wobei nur vier Texte einen Wert von $C < 0.02$ aufweisen und das Gesamtkorpus mit dieser Verteilung überhaupt nicht modelliert werden

kann. Insofern ist diese Verteilung somit als ungeeignet für die Modellierung ukrainischer Graphemfrequenzen anzusehen.

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 $C = 0.0244$. Interessanterweise tendiert hierbei (wie auch in einer ganzen Reihe der einzelnen Texte) bei einem Teil der Texte der Parameter a gegen 0, bei dem anderen Teil der Texte der Parameter p gegen 1 – was insofern von Bedeutung ist, als sich für $a = 0$ die (1-verschobene) geometrische Verteilung und für $p = 1$ die Zeta-Verteilung als ein Spezialfall der Good-Verteilung erweist (Wimmer/ Altmann 1999: 219f.).

Tabelle 4a/b

Anpassung der rechts-gestutzten Zeta-Verteilung und der Zipf-Mandelbrot-Verteilung an 30 ukrainische Texte

	rechts-gestutzt Zeta, R=33			**Zipf-Mandelbrot (a,b) n=33**			
Nr.	***a***	***Chi²$_{FG=30}$***	***C***	***a***	***b***	***Chi²$_{FG=29}$***	***C***
1	0,66	627,64	0,1105	1,18	3,51	360,91	0,0635
2	0,67	268,44	0,0879	1,63	7,45	101,47	0,0332
3	0,63	2495,60	0,1026	1,16	3,92	1433,91	0,0590
4	0,64	869,30	0,1231	6,23	56,58	169,13	0,0239
5	0,64	804,31	0,1262	6,28	55,28	135,55	0,0213
6	0,63	693,57	0,1221	2,60	17,32	175,15	0,0308
7	0,67	869,05	0,1166	4,07	30,70	166,25	0,0223
8	0,66	793,86	0,1149	5,45	45,90	154,59	0,0224
9	0,66	760,03	0,1195	12,00	117,52	80,35	0,0126
10	0,60	796,51	0,1255	1,19	4,37	472,09	0,0744
11	0,62	1164,69	0,1287	1,36	5,53	565,47	0,0625
12	0,61	967,61	0,1409	5,76	52,08	239,18	0,0348
13	0,61	910,22	0,0968	3,35	28,29	266,34	0,0283
14	0,63	890,13	0,1324	1,26	4,40	507,28	0,0754
15	0,64	2138,16	0,1256	1,19	3,69	1283,04	0,0753
16	0,63	1512,62	0,0963	2,51	17,78	476,76	0,0304
17	0,63	2428,63	0,1354	4,01	31,34	706,73	0,0394
18	0,64	3987,79	0,1042	1,17	3,74	2296,50	0,0600
19	0,64	927,55	0,1232	1,73	8,59	438,24	0,0582
20	0,64	2481,30	0,1123	1,17	3,72	1478,96	0,0669

* In zwei Fällen war eine Anpassung der Good-Verteilung überhaupt nicht möglich (Texte #1 und #3).

21	0,65	1612,81	0,1076	1,20	3,80	861,17	0,0575
22	0,65	1434,28	0,1057	2,43	15,53	422,65	0,0311
23	0,66	1784,42	0,1053	1,20	3,64	952,78	0,0562
24	0,65	2612,36	0,1077	1,21	3,84	1459,19	0,0602
25	0,66	2734,65	0,0984	1,15	3,28	1550,73	0,0558
26	0,64	2480,07	0,0951	1,17	3,77	1352,21	0,0518
27	0,67	980,57	0,1245	1,84	8,70	381,20	0,0484
28	0,64	1037,31	0,1259	2,72	17,95	327,57	0,0397
29	0,65	644,18	0,1174	3,63	27,24	152,52	0,0278
30	0,67	1215,41	0,1057	7,38	66,21	225,49	0,0196

Tabelle 5a/b

Anpassung der rechts-gestutzten geometrischen Verteilung und der Good-Verteilung an 30 ukrainische Texte

	rechts-gestutzt geo-metrisch (*q*, *R* = 33)			**Good-1 (*a*, *p*)**			
Nr.	**q**	***Chi²$_{FG=30}$***	**C**	**a**	**p**	***Chi²$_{FG=30}$***	**C**
1	0,9129	156,59	0,0276	-	-	-	-
2	0,9142	77,90	0,0255	0,0000	0,9055	77,41	0,0680
3	0,9186	696,38	0,0286	-	-	-	-
4	0,9150	162,93	0,0231	0,7169	0,9800	900,97	0,3779
5	0,9133	115,64	0,0181	0,7175	0,9800	822,10	0,3790
6	0,9155	93,39	0,0164	0,0000	0,9061	95,11	0,0599
7	0,9104	124,11	0,0167	0,7348	0,9800	876,42	0,3554
8	0,9118	143,06	0,0207	0,7298	0,9800	811,49	0,3580
9	0,9127	95,00	0,0149	0,0000	0,9041	93,47	0,0546
10	0,9195	198,12	0,0312	0,0000	0,9089	207,23	0,0815
11	0,9172	189,18	0,0209	0,0000	0,9072	196,28	0,0671
12	0,9172	197,40	0,0287	0,0000	0,9069	204,81	0,0751
13	0,9215	249,74	0,0266	0,6867	0,9800	1031,72	0,3744
14	0,9149	196,54	0,0292	0,7058	0,9800	926,71	0,3972
15	0,9128	486,24	0,0286	0,7160	0,9800	2359,66	0,1386
16	0,9188	381,27	0,0243	0,6989	0,9800	1677,42	0,3633
17	0,9135	513,29	0,0286	0,7112	0,9800	2505,48	0,3962
18	0,9168	1075,68	0,0281	0,7057	0,9800	14048,95	0,1215
19	0,9135	236,61	0,0314	0,0000	0,9043	239,45	0,0724
20	0,9154	632,27	0,0286	0,0000	0,9059	644,49	0,0725
21	0,9150	328,57	0,0219	0,7151	0,9800	1707,37	0,3624
22	0,9146	261,12	0,0192	0,7156	0,9800	1514,09	0,3591
23	0,9136	353,77	0,0209	0,0000	0,9047	353,01	0,0620
24	0,9140	654,77	0,027	0,0000	0,9051	654,86	0,0690
25	0,9139	752,67	0,0271	0,7242	0,9800	2923,34	0,3463
26	0,9178	671,29	0,0257	0,0000	0,9082	759,59	0,0291

27	0,9098	162,83	0,0207	0,0000	0,9015	159,25	0,0566
28	0,9130	198,07	0,024	0,7162	0,9800	1065,84	0,3803
29	0,9128	106,89	0,0195	0,7207	0,9800	664,22	0,3676
30	0,9113	235,71	0,0205	0,7327	0,9800	1253,55	0,3462

Damit können wir die ersten vier der sechs von uns betrachteten – und hierbei handelt es sich um die in der bisherigen Forschung am häufigsten diskutierten – Verteilungsmodelle mitsamt als ungeeignet für die Modellierung der Ranghäufigkeit ukrainischer 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. Im Hinblick auf rangierte Graphemhäufigkeiten hatte sie erstmals Grzybek (2001) auf der Basis eines Textes von A.S. Puškin („Царь Салтан“) ins Spiel gebracht, der mit einem Wert von $C = 0.0082$ auf ein ausgezeichnetes Anpassungsergebnis kam.

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; ähnliches auch für Slowakisch (vgl. Grzybek/Kelih/Altmann 2005a, b) und für das Slowenische (vgl. Grzybek/Kelih/Altmann 2005c). 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 ukrainischen Texte ergibt: die sich aufgrund der angeführten Parameter ergebenden theoretischen Häufigkeiten NP_i sind neben den beobachteten Häufigkeiten f_i der 33 ukrainischen 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 sehr gutes Modell dar ($C = 0.0076$).

Tabelle 6

Beobachtete und theoretische Werte (negativ hypergeometrische Verteilung) für das Gesamtkorpus

i	f_i	NP_i	***i***	f_i	NP_i
1	37267	38517,34	**18**	8877	9032,86
2	32774	30204,66	**19**	7487	8348,38
3	25080	26434,40	**20**	6888	7683,92
4	24639	23917,34	**21**	6406	7037,85
5	21053	21979,66	**22**	5850	6408,83
6	20941	20374,66	**23**	5074	5795,74
7	20075	18985,49	**24**	4625	5197,65
8	19171	17747,87	**25**	3876	4613,82
9	16296	16622,74	**26**	3843	4043,68
10	16240	15584,62	**27**	3565	3486,83
11	13936	14615,97	**28**	2857	2943,04
12	13697	13704,23	**29**	2790	2412,33
13	12959	12840,11	**30**	2484	1895,00
14	12949	12016,51	**31**	2407	1391,85
15	12398	11227,94	**32**	506	904,54
16	10584	10470,04	**33**	78	436,78

17	8944	9739,30			
	K =2,8688			χ^2 =2926,52	
	M =0,8101			FG =29	
	n =32			C =0,0076	

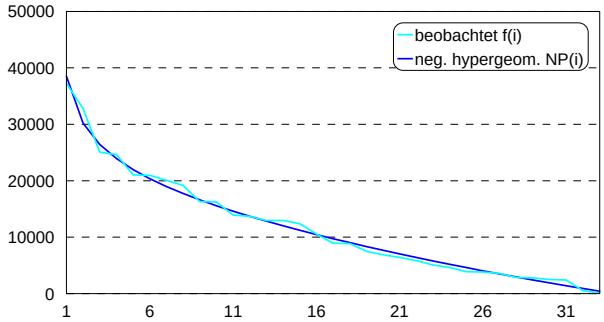

Abb. 3
Anpassung der negativen hypergeometrischen Verteilung an 30 ukrainische Texte

Tab. 7 zeigt die Ergebnisse der Anpassungen für die 30 Einzeltexte; es bestätigen sich auch hier die hervorragenden Ergebnisse: Der Diskrepanzkoeffizient liegt insgesamt im Intervall von $0{,}0145 \geq C \geq 0.0032$; dabei haben alle 30 Einzelanalysen einen Wert von $C < 0.02$, wobei 20 Texte sogar einen Wert von $C < 0.01$ aufweisen, was insgesamt als ein hervorragendes Ergebnis zu interpretieren ist. Abb. 4 veranschaulicht die Werte des Diskrepanzkoeffizienten C für die 30 ukrainischen Texte.

Tabelle 7
Ergebnis der Anpassung der negativ hypergeometrischen Verteilung an 30 ukrainische Texte

Neg. Hypergeometrisch					**Neg. Hypergeometrisch**				
Nr.	**K**	**M**	***Chi²$_{FG=29}$***	**C**	**Nr.**	**K**	**M**	***Chi²$_{FG=29}$***	**C**
1	2,932	0,801	55,34	0,01	**16**	2,816	0,801	104,98	0,007
2	2,742	0,752	33,66	0,011	**17**	3,212	0,872	78,99	0,004
3	2,851	0,812	221,35	0,009	**18**	2,888	0,808	261,2	0,007
4	3,05	0,844	87,63	0,012	**19**	3,099	0,842	37,45	0,005
5	3,021	0,83	64,12	0,01	**20**	2,955	0,82	141,81	0,006
6	2,932	0,827	82,48	0,015	**21**	2,911	0,816	115,65	0,008
7	2,901	0,793	94,07	0,013	**22**	2,902	0,812	62,26	0,005
8	2,948	0,805	64,73	0,009	**23**	2,931	0,803	93,64	0,006
9	2,976	0,826	77,76	0,012	**24**	2,88	0,798	258,41	0,011

10	3,09	0,88	34,28	0,005	**25**	2,827	0,774	295,06	0,011
11	3,07	0,865	82,05	0,009	**26**	2,769	0,782	268,93	0,01
12	3,238	0,905	37,86	0,006	**27**	3,052	0,813	60,7	0,008
13	2,787	0,807	95,42	0,01	**28**	2,986	0,818	89,72	0,011
14	3,122	0,86	56,6	0,008	**29**	3,043	0,829	17,51	0,003
15	3,031	0,825	132,37	0,008	**30**	2,929	0,789	61,89	0,005

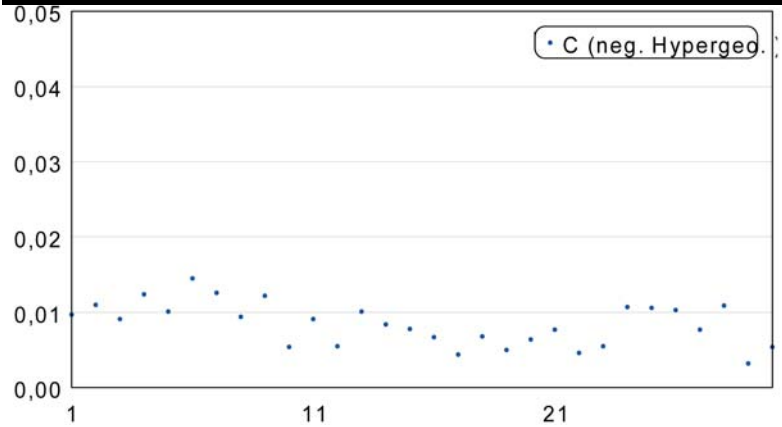

Abb. 4:
Diskrepanzkoeffizient *C* (negative hypergeometrische Verteilung) für 30 ukrainische Texte

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; vor allem auch die Parameter *K* und *M* stellen sich als ziemlich konstant dar: Abgesehen von dem ohnehin konstanten Parameter *n* (der konstant bei $n = 32$, d.h. um eins niedriger als der Inventarumfang liegt), liegen die Werte für *K* im Intervall von $3.24 \geq K \geq 2.74$, wobei das 95%-Konfidenzintervall bei einer Unter- bzw. Obergrenze von $K_u = 2.92$ und $K_o = 3.00$ relativ eng ist. Dasselbe gilt für den Parameter *M*, der im Intervall zwischen und $0.91 \geq M \geq 0.75$ liegt, wobei auch hier das 95%-Konfidenzintervall bei einer Unter- bzw. Obergrenze von $M_u = 0.81$ und $M_o = 0.83$ sehr schmal ist. Abb. 5 veranschaulicht die relative Konstanz der Ergebnisse.

In Anbetracht der Tatsache, dass die negativ hypergeometrische Verteilung drei Parameter aufweist, von denen einer direkt vom Inventarumfang abhängt, ist abschließend von Interesse, ob sich auch die Whitworth-Verteilung für ukrainische 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 ukrainischen Texten im Detail dar. Deutlich ist zu sehen, dass die Whitworth-Verteilung sich nicht für die Modellierung der ukrainischen Daten eignet: Der Diskrepanzkoeffizient liegt für die einzelnen Stichproben im Intervall $0.1207 \geq C \geq 0.0210$. Damit scheidet die Whitworth-Verteilung für die Modellierung als geeignetes Modell aus.

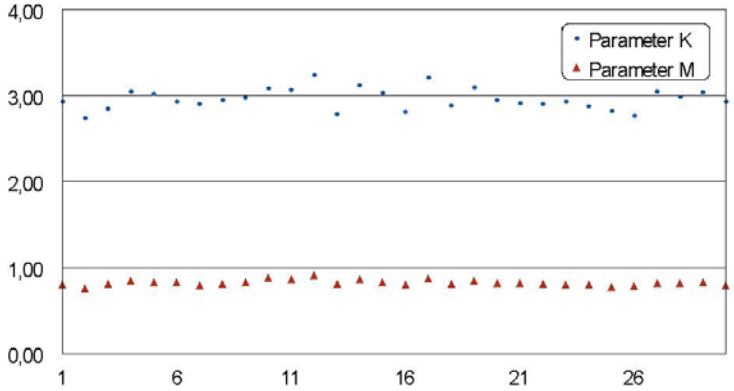

Abb. 5:
Konstanz der Parameter *K* und *M*

Tabelle 8
Ergebnis der Anpassung der Whitworth-Verteilung an 30 ukrainische Texte

W h i t w o r t h, R = 33					
Nr.	***Chi²$_{FG=31}$***	**C**	**Nr.**	***Chi²$_{FG=31}$***	**C**
1	119,08	0,0210	**16**	1627,71	0,1036
2	70,30	0,0230	**17**	2164,41	0,1207
3	2614,52	0,1075	**18**	4153,54	0,1085
4	815,63	0,1155	**19**	886,99	0,1178
5	729,23	0,1144	**20**	2471,35	0,1118
6	633,88	0,1116	**21**	1635,40	0,1092
7	849,05	0,1139	**22**	1456,26	0,1073
8	784,96	0,1136	**23**	1846,30	0,1090
9	719,69	0,1131	**24**	2707,87	0,1117
10	730,26	0,1151	**25**	3041,65	0,1094

11	1031,59	0,1140	**26**	2725,74	0,1045
12	828,27	0,1206	**27**	929,14	0,1180
13	977,50	0,1039	**28**	955,93	0,1160
14	799,95	0,1190	**29**	616,22	0,1123
15	2008,85	0,1180	**30**	1275,00	0,1109

ZUSAMMENFASSUNG, SCHLUSSFOLGERUNGEN UND PERSPEKTIVEN

Aufgrund der in der vorliegenden Untersuchung angestellten theoretischen Überlegungen und empirischen Befunde ergibt 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 ukrainischer Grapheme nicht eignen; dieser Befund deckt sich mit den Ergebnissen zum Russischen (Grzybek/Kelih/Altmann 2004), Slowakischen (Grzybek/Kelih Altmann 2005a,b) und Slowenischen (Grzybek/Kelih 2003). 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. Auch die Whitworth-Verteilung führt im Falle des Ukrainischen zu keinen befriedigenden Resultaten; dies ist ein klarer Gegensatz im Vergleich zu den Befunden zum Russischen, deckt sich aber mit den Ergebnissen zum Slowakischen und Slowenischen.
3. Als einziges und ausgezeichnetes Modell für die Modellierung rangierter Graphemhäufigkeiten des Ukrainischen eignet sich die negativ hypergeometrische Verteilung, mit der sich hervorragende Anpassungsergebnisse erzielen lassen. Allerdings hat diese Verteilung nicht weniger als drei Parameter (n, K, M), von denen nur einer (n) sich direkt als vom Inventarumfang abhängig interpretieren lässt.
4. Interessant sind die Hinweise auf eine mögliche Interpretation der Parameter K und M der negativ hypergeometrischen Verteilung: Vergleicht man nämlich die über die 30 ukrainischen Texte hinweg die augenscheinlich relativ konstanten Werte für die Parameter K und M der negativ hypergeometrischen Verteilung mit den

Werten, die diese Parameter im Slowenischen, Russischen und Slowakischen einnehmen, so stellt sich heraus, dass die Parameterwerte von *K* für das Slowenische (mit 25 Graphemen) deutlich niedriger und für das Slowakische (mit 43 bzw. 46 Graphemen) deutlich höher liegen als für das Ukrainische und Russische, während sich die Parameterwerte von *M* anscheinend nicht wesentlich von denen im Russischen unterscheiden. Dieser Umstand ist gegebenenfalls als Indiz dafür zu werten, dass die Parameter *K* und *M* in zumindest indirekter Abhängigkeit vom Inventarumfang zu interpretieren sein könnten. Eine Überprüfung dieser Annahme kann jedoch nur in vergleichender Analyse – möglichst mit weiteren empirischen Studien an Sprachen mit unterschiedlichem Umfang des Grapheminventars vorgenommen werden (Grzybek/Kelih/Altmann 2005a,b,c). Vor diesem Hintergrund wird dann auch die Frage der (fehlenden) Berücksichtigung des Apostrophs im Ukrainischen neu zu reflektieren sein.

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

1. Wie die Analyse der ukrainischen Texte gezeigt hat, scheinen schlechte Anpassungsergebnisse insbesondere bei kürzeren Texten vorzukommen; dies macht es erforderlich, dem Faktor der (notwendigen bzw. optimalen) Stichprobengröße in Zukunft systematisch nachzugehen (vgl. Grzybek/Kelih/Altmann 2005d).
2. 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 2005d).

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.

LITERATUR

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. and Kelih, E. (2005): Textforschung: Empirisch! In: J. Banke, B. Dumont (Eds.): Textsortenforschungen. Leipzig. FSR, 2005, 13-34.

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 Slowakischen (Teil I: Ohne Digraphen). In Nemcová, E., ed.: Philologia actualis slovaca. UCM, Trnava. [in Druck]

Grzybek, P.; Kelih, E.; Altmann, G. (2005b): Graphemhäufigkeiten im Slowakischen (Teil II: Mit Digraphen). In: Sprache und Sprachen in Mitteleuropa. GeSuS, Trnava (2005). [in Druck]

Grzybek, P.; Kelih, E.; Altmann, G. (2005c): „Graphemhäufigkeiten im Slowenischen.“ In: *Slavistična Revija*. [In print]

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

Grzybek, P.; Kelih, E.; Stadlober, E. (2005): Empirische Textsemiotik und quantitative Texttypologie. In: Bernard, J.; Fikfak, Ju.; Grzybek, P. (Eds.): *Text & Reality*. Ljubljana/Wien/Graz: ZRC, 95-120.

Grzybek, P., Stadlober, E., Kelih, E., and Antić, G. (2005): Quantitative Text Typology: The Impact of Word Length. In: C. Weihs and W. Gaul (Eds.), Classification – The Ubiquitous Challenge. Springer, Heidelberg; 53-64.

Kelih, E., Antić, G., Grzybek, P. and Stadlober, E. (2005) Classification of Author and/or Genre? The Impact of Word Length. In: C. Weihs and W. Gaul (Eds.), Classification – The Ubiquitous Challenge. Springer, Heidelberg, 498–505.

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.

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

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.; 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.

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.