What a Difference an »E« Makes: Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen

Peter Grzybek (Graz)

Wenn in gegenwärtigen Diskursen vereinfachend von „dem“ Roman ohne »E« die Rede ist, so bezieht sich das in der Regel auf den ursprünglich 1969 erschienenen Roman *La Disparition* von George Perec. Dieser original in französischer Sprache verfasste Text ist in der Tat durch das ihn auszeichnende Merkmal charakterisiert, dass der Buchstabe »E« im gesamten Text nicht ein einziges Mal vorkommt. Doch trotz seiner Originalität ordnet der Roman sich natürlich in eine bestimmte Tradition ein, in der sich nicht minder berühmte Vorläufer ausmachen lassen. Einer der bekanntesten von diesen ist sicherlich der amerikanische Roman *Gadsby. A Story of over 50000 words withouth using the letter E* von Ernest Vincent Wright aus dem Jahre 1939, zu dem in den Roman von Perec deutliche intertextuelle Bezüge eingebaut sind. Freilich lässt sich die hinter dieser Schreibe liegende Tradition viel weiter, bis in die Antike und ins Mittelalter, zurückverfolgen und durch die ganze Literatur- und Kulturgeschichte verfolgen. Das kann und muss an dieser Stelle nicht im Detail getan werden, zumal niemand anders als George Perec selbst dies ausführlich getan hat, und zwar 1973 in seinem Aufsatz „Histoire de la lipogramme“.

Die Bezeichnung ‚Lipogramm’ geht zurück auf das griechische λέιπειν (leipein) für *weglassen* und γραμμα (gramma) für Buchstabe; Lipogramme bzw. Leipogramme sind folglich Texte, in denen auf die Verwendung eines oder mehrerer Buchstaben des Alphabets verzichtet wird. Insbesondere in der Blütezeit dieser literarischen Tradition im spanischen und deutschen Barock erschienen immer wieder entweder vollständige Erzählungen oder aber in größere Texte eingeflochtene Passagen ohne bestimmte Buchstaben. Erwähnt seien nur die *Frauenzimmer Gesprächsspiele* (1644-57) von Georg Philipp Harsdörffer (1607-1658), in denen Erzählungen ohne »M« und »L« vorkommen, oder der Roman *Die drei ärgsten Erznarren der ganzen Welt* (1672) von Christian Weise (1642-1708), mit einer R-losen Rede für einen verliebten Mann, der das »R« nicht aussprechen konnte. Im 19. Jahrhundert ist es im Französischen vor allem der Roman *Curieux voyage autour du monde* (1853) von Jacques Aragos (1790-

1855); im Deutschen ist insbesondere Franz Rittler bekannt mit seinem 1815 erschienenen Roman *Die Zwillinge. Versuch, aus sechzig aufgegebenen Worten einen Roman ohne R zu schreiben* bekannt, der nur wenig später von seinem lipogrammatischen Mitstreiter Leopold Kolb mit dessen Roman ohne »R« *Keine Liebe ohne Qualen. Eine kleine Geschichte, einfach und doch künstlich* beantwortet wurde. Ähnlich wie diesen beiden an einem Nachweis der Reichhaltigkeit und Biegsamkeit der deutschen Sprache lag, wollte nahezu zeitgleich im Russischen niemand anders als Gavrila Romanovič Deržavin (1743-1816) den Nachweis der Flexibilität und Weichheit der russischen Sprache („dokazatel'stvo izobilija i mjagkosti russkogo jazyka") erbringen, man vergleiche nur in seinen Anakreontische Liedern zu findenden Verstexte ohne »R« wie „Solovej vo sne" (1797) oder „Šutočnoe želanie" (1802). Im 20. Jahrhundert schließlich haben solche Texte stärker experimentellen Charakter, so dass der Bogen möglicher Hintergründe dieser Art zu schreiben weitgehend abgesteckt ist: angefangen von buchstabenmystischen Konzeptionen reicht dieser über avantgardistische und experimentelle Poesie bis hin zum reinen Sprachspiel.

Das Grundkonzept von *La Disparition* ist vor diesem Hintergrund ein wenig anders und anspruchsvoller: Auf den ersten Blick scheint dieses vergleichsweise einfach, insofern das Fehlen des häufigsten der 26 Buchstaben des französischen Alphabets auf der formalen Ebene auf inhaltlicher Ebene durch das Verschwinden einer Figur gespiegelt wird: Allein der Name des Protagonisten, Anton Voyl – der im übrigen auf den unbetonten Vokal »E« [voyelle atone] verweist –, ist hier Programm. Doch in den Text sind auf verschiedenste Art und Weise weitere zahlreiche Spuren zum Buchstaben E und um ihn herum gelegt; neben vielen feinsinnigen Anspielungen wirkt sich das bis auf die makrostrukturelle Organisation des Textes hin aus: So lässt sich die Anzahl der sechs Teile des Romans als Entsprechung zu den sechs Vokalgraphemen A, E, I, O, U, Y sehen; auch die Tatsache, dass der Roman aus insgesamt 26 Kapiteln besteht, ist bei der Anzahl von 26 Buchstaben des französischen Alphabets nur schwer als Zufall abzutun. Zusätzliche Nahrung erhalten derartige Parallelen durch den Umstand, dass das fünfte Kapitel des Romans ebenso fehlt wie die Überschrift zum zweiten Teil, was vor dem Hintergrund, dass es sich beim »E« nicht nur um den zweiten von fünf Vokalen, sondern auch um den fünften Buchstaben des Alphabets handelt, ein insgesamt konsistentes Bild ergibt. Literaturwissenschaftliche Untersuchungen haben mittlerweile den literarischen Hintergrund des Werkes von Perec ausführlich beleuchtet. Dabei wurde der Roman in den Kontext der 1960 gegründeten literarischen Gruppe *Oulipo* eingebettet, deren Namen auf die Bezeichnung

***Ou**vroir de **li**ttérature **po**tentielle* zurückgeht; mit ihrem Konzept der *écriture sous constraintes*, dem Schreiben unter Restriktionen, verweist diese Gruppe auf den für sie essentiellen Zusammenhang zwischen Literatur und Mathematik (vgl. Schleypen 2004): Die Grundidee dieser personellen und konzeptuellen Verbindung geht auf die Überlegung zurück, dass Literatur im Grunde genommen durch formale Regeln geprägt sei, denen jegliches (literarische) Schreiben unterworfen sei. Diese seien mathematischen Strukturen vergleichbar, aus denen jeweils spezifische (durchaus dem Jakobson'schen Poetizitätsmodell analoge) Restriktionsbedingungen als Quelle poetischer Kreativität hervorgehen.

Während die Frage von Restriktionen und deren Folgen für das Schreiben also in der theoretischen Reflektion durchaus prominenten Raum einnimmt, haben bisherige Forschungen zu konkreten Auswirkungen derartiger Restriktionen auf die sprachliche Gestaltung eher Seltenheitswert. Insbesondere den genuin sprachlichen Auswirkungen der von Perec in seinem Roman verfolgten und befolgten paradigmatischen Reduktion des Buchstabeninventars ist in der bisherigen Forschung nie systematisch nachgegangen worden, so dass diese eigentlich elementare Frage ein echtes Forschungsdesiderat darstellt. Eine seltene Ausnahme hierzu stellt in gewisser Hinsicht die Untersuchung „Lorsque l'E disparaît du cercle des voyelles" von Chesneau (1984): Er stellt auf der Grundlage der in den Roman von Perec integrierten Nachdichtung von Baudelaires „Recuiellement" an verschiedenem sprachlichen Material phonetischer, silbischer, metrischer und morphologischer Art (so etwa der Anteil offener und geschlossener Silben) einen Vergleich zum Original bzw. zur vermeintlichen Norm der französischen Sprache an und arbeitet als wesentliches Ergebnis heraus, dass Nachdichtung und Original sich – abgesehen von der Verteilung maskuliner und femininer Substantive – strukturell nicht wesentlich voneinander unterscheiden.

Die im Grunde genommen viel essentiellere, und scheinbar auch viel „einfachere" Frage, wie sich das Fehlen des häufigsten Vokals und Buchstaben auf die eigentliche Struktur der Buchstabenhäufigkeiten auswirkt, ist bislang jedoch merkwürdigerweise nie gestellt worden. Grund dafür dürfte sein, dass eine solche Fragestellung nicht nur im Grenzbereich zwischen Sprach- und Literaturwissenschaft liegt, sondern zudem die etablierten Grenzen beider Disziplinen überschreitet, indem sie die Spielwiese der hermeneutischen Beschäftigung verlässt, den ohne Bezugnahme auf mathematische und statistische Konzepte nicht beschreibbaren Raum wissenschaftlicher Textuntersuchungen betritt, und diesen Raum erst danach wieder durch die Tür verlässt, hinter der sich die Option einer qualitativen Beschreibung befindet.

Wie also verhält es sich – so lautet die Gretchenfrage in einfacherer Form – mit den Häufigkeiten der einzelnen Buchstaben im Text? Was ändert sich an der Gesamtheit eines Systems, was ändert sich an den Prinzipien und Prozessen der Systemregulierung, wenn ein System seines ansonsten am häufigsten vorkommenden Elements beraubt wird? Eine Beantwortung dieser Frage weist weit über triviale Buchstabenzählungen hinaus in die Tiefen systemtheoretischer Analyse und hat eine mehr oder weniger ausreichende Kenntnis des Systemverhaltens von Buchstabenhäufigkeiten zur Voraussetzung. Diese Kenntnis ist jedoch im Hinblick auf das französische und die Graphemhäufigkeiten des französischen Alphabets schlicht und einfach nicht gegeben. Zwar gibt es gerade zum Französischen schon seit der ersten Hälfte des 19. Jahrhunderts Untersuchungen zur Buchstabenhäufigkeit im Französischen, die ihre spätere Fortsetzung in Arbeiten wie denen von Galland (1941) haben und sich heutzutage in verschiedenen Internetquellen manifestieren, die mittlerweile sogar scripts zur automatischen Berechnung von Häufigkeiten beinhalten.[1] Doch all diese Quellen sind mit einer ganzen Reihe von Problemen verbunden, die eine systematische Analyse bislang nicht erlauben; ein wesentliches Grundproblem liegt allein schon darin begründet, dass die meisten Quellen von unterschiedlichen Inventarumfängen ausgehen: So basiert etwa die Zählung von http://pedroiy.free.fr/alphabets/frequence.htm auf den 26 »core letters«, während die in der französischen Wikipedia angeführten Zählungen[2] unter Berücksichtigung verschiedener (allerdings nicht aller!) diakritischer Zeichen auf 36 Buchstaben beruhen.

Noch problematischer aber ist, dass bislang eigentlich keinerlei Ansätze vorliegen, die Vorkommenshäufigkeiten der Buchstaben des französischen Alphabets in einen theoretischen Kontext zu stellen, d.h. die Vorkommenshäufigkeiten selbst als Ergebnis eines selbst-regulierenden Prozesses theoretisch zu beschreiben und zu modellieren. Im Grunde genommen richten sich bisherige Studien ausschließlich auf die Frage nach der Vorkommenshäufigkeit jeweils individueller Buchstaben und Grapheme. Derartige Häufigkeitsverteilungen können im Rahmen von systematischen und system-bezogenen Untersuchungen allerdings nicht mehr und nicht weniger als Rohmaterial und Ausgangspunkt für weiterführende Fragestellungen sein: Denn unter system-bezogener bzw. systemtheoretischer Perspektive geht es nicht um das spezifische Verhalten individueller Elemente in ver-

1 http://euler.ac-versailles.fr/webMathematica/versailles/stats/frequences_lettres/frequences_lettres.jsp [Stand 8.6.2007]

2 http://fr.wikipedia.org/wiki/Fr%C3%A9quence_d‘apparition_des_lettres_en_fran%C3%A7ais [Stand 8.6.2007]

schiedenen Stichproben (bildlich in den Worten Tynjanovs gesprochen: um deren „Auto-Funktion“), sondern um deren systemisches Verhalten im relationalen und funktionalen Kontext des gegebenen Systems (ihre „Syn-Funktion“) .
Diese Perspektive verändert den Blickwinkel auf das Material nachhaltig: Denn anstelle eines Fokus auf individuelle Buchstaben geht es um das Verhalten der Buchstaben in der Gesamtheit des Systems, im gegebenen Fall also um die durch die Restriktion bedingte Veränderung des Systemverhaltens. Dieser Frage soll in den folgenden Analysen detaillierter nachgegangen werden:

> Wie lässt sich das häufigkeitsbezogene Systemverhalten der Buchstaben durch Bezugnahme auf ein theoretisches Häufigkeitsmodell beschreiben, und welche Modifikationen ergeben sich unter der Bedingung des Ein- oder Ausschlusses der genannten Restriktion auf das globale Systemverhalten der Graphemhäufigkeiten?

Diese Fragen sind durchaus im Sinne der Oulipo-Konzeption zu verstehen, insofern sie einerseits auf die sprachliche Spezifik der Einführung von Restriktionen zielen, andererseits genau damit aber auch Einsichten in das „normale“ Systemverhalten und die Möglichkeit von dessen systematischer Beschreibung versprechen. Untersucht werden sollen diese Fragen jedoch nicht am französischen Originaltext, sondern vielmehr am Text der deutschen Übersetzung durch Eugen Henlé. Der Grund dafür ist das bereits angesprochene Fehlen systematischer theoriebezogener Untersuchungen zu Graphemhäufigkeiten des Französischen unter „normalen“ Bedingungen, die hier nicht „en passant“ nachgeliefert werden können. Dasselbe gilt für die englische Übersetzung *A Void*, die Gilbert Adair 1994 angefertigt hat. Eher anbieten würden sich aufgrund der Forschungslage eine Untersuchung zum Russischen: Denn zum einen sind in jüngster Zeit eine Reihe systematischer Untersuchungen zum Russischen erschienen (Grzybek/Kelih 2003; Grzybek/Kelih/Altmann 2004, 2005a; Grzybek 2005; Grzybek/Kelih 2005), in denen das Häufigkeitsverhalten russischer Grapheme detailliert analysiert und im Kontext anderer slawischer Sprachen in ein allgemeines theoretisches Modell eingeordnet worden ist; zum anderen steht seit 2005 mit *Isčezanie* eine russische Übersetzung als Vergleichbasis zur Verfügung, auch wenn hier das Fehlen des »O«, und nicht das »E« als häufigster Buchstabe des Russischen die Restriktionsbedingung darstellt (vgl. Kislov 2006).
Dennoch aber soll im hier gegebenen Rahmen einer Analyse des Deutschen der Vorzug gegeben werden, was spätere Vergleiche zu anderen Sprachen ja nicht ausschließt. Denn auch zu deutschen Graphemhäufigkeiten liegen mit der Untersuchung von Best (2005) erste Versuche vor, auf Regularitäten im Häufigkeitsverhalten hinzuweisen. Diese Untersuchungen richten sich freilich nicht nur auf Buchstaben; vielmehr enthalten sie mitunter auch das Vorkommen anderer Schriftzeichen wie etwa Leerzei-

chen, Apostroph, Gedankenstrich u.a.m., die in die allgemeinen Häufigkeiten eingehen. Damit verändern sich natürlich einerseits die relativen Häufigkeiten der eigentlichen Grapheme, andererseits auch der jeweils zugrunde liegende Inventarumfang. Tab. 1 enthält die entsprechenden Textangaben (*N* ist der Stichprobenumfang, d.h. die Graphemanzahl pro Text, *I* ist der jeweils zugrunde gelegte Inventarumfang).

Tab. 1: Stichproben aus Best (2005)

Nr.	**Text**	*N*	*I*	**Anmerkungen zum Buchstabeninventar**
1	H. Pestalozzi: Hühner, Adler und Mäuse	675	27	Q,X,Y kommen nicht vor
2	G.A. Bürger: Münchhausen	137476	30	
3	G.A. Bürger: Lenore	6215	27	Q,X,Y kommen nicht vor
4	G. Büchner: Lenz	42608	30	
5	G. Büchner: Hessischer Landbote	21452	30	
6	K. May: Winnetou I	777361	32	É, Ñ; Ae → Ä, Oe → Ö, Ue → Ü
7	F. Kafka: Die Verwandlung	99559	30	
8	F. Kafka: Der Prozeß	361848	30	
9	G. Vesper: Fugen	6259	27	Q,X,Y kommen nicht vor
10	O. Jägersberg: Dazugehören	40977	30	
11	J. Joffe: Nach dem Bruderkrieg	6091	30	
12	R. Hoppe: Das gierige Gehirn	20075	30	
13	Schönpflug (1969)	99984	29	ohne: ß
14	K.H. Best: Wiss. Prosa	179922	30	

Da jedoch gerade der Inventarumfang für die Modellierung der Häufigkeiten einer der entscheidenden Faktoren zu sein scheint – wie in den o.a. Analysen zu slawischen Sprachen gezeigt wurde – hat Grzybek (2007) die Daten von Best einer Re-Analyse unterzogen, in denen das Datenmaterial ausschließlich auf die eigentlichen Grapheme reduziert wurde. Das auf diese Art und Weise „modifizierte“ Datenmaterial verhielt sich im Hinblick auf die Buchstabenhäufigkeiten und deren Modellierung überaus systematisch. Aus diesem Grunde soll es im vorliegenden Text als Ausgangsbasis für einen Vergleich mit der von Perec eingeführten Restriktionsbedingung dienen. Damit liegen zunächst zwei „Versuchsbedingungen“ vor: Die durch den Text von Perec repräsentierte Restriktionsversion soll im folgenden als **D-R** bezeichnet werden. Die Basisversion hingegen soll als **D-3**0 bezeichnet werden: Sie beruht auf den wie oben beschrieben „mo-

difizierten“ Daten von Best; dabei wird ein Inventarumfang von 30 Graphemen unter Einschluss der 26 Standardbuchstaben des Deutschen, der drei Umlaute Ä, Ö, Ü und des nur in Kleinschreibung verwendeten ß zugrunde gelegt.[3] Ergänzt werden soll dieses Bedingungspaar durch eine dritte Vergleichsoption, die darin besteht, aus den berechneten Häufigkeiten der Bedingung D-30 ***im nachhinein*** die jeweils größte Häufigkeit – und dies ist in den betreffenden Stichproben durchgehend der Buchstabe »E« – zu eliminieren, so dass ein Inventarumfang von 29 Elementen übrig bleibt. Diese Bedingung, die im folgenden als **D-29** bezeichnet wird, erlaubt somit einen zweifachen Vergleich mit D-30 und D-R. Zur Untersuchung der Bedingung D-R wurden die ersten 10 Kapitel (incl. des Vorworts) analysiert (und dabei zwei beim Druck fälschlicherweise in den Text geratene »E« eliminiert!); jede dieser Stichproben ist ausreichend groß und insofern aussagekräftig genug, wobei sich die Analyse mehrerer Stichproben (und nicht etwa einer Korpusanalyse) als sinnvoll erwiesen hat, um die „natürliche“ Varianz innerhalb einer Sprache abzudecken. Tab. 2 enthält die Umfänge der einzelnen Stichproben.

Tab. 2: Stichproben aus der deutschen Perec-Übersetzung (D-R)

Nr.	**Kapitel**	***N***	Nr.	**Kapitel**	***N***
1	Vorwort	5687	6	Kap. V	15306
2	Kap. I	12141	7	Kap. VI	19840
3	Kap. II	19318	8	Kap. VII	14371
4	Kap. III	15305	9	Kap. VIII	11945
5	Kap. IV	5154	10	Kap. IX	17034

Beginnen wir die Untersuchung also mit einem Vergleich des globalen Systemverhaltens deutscher Buchstaben unter den genannten drei Bedingungen.

Das globale Systemverhalten der Buchstaben: Entropie und Redundanz

Die statistische Eigenschaft eines Systems ist dann als global anzusehen, wenn sie sich nicht auf einzelne Entitäten, sondern auf das ganze Ensemble bezieht. Bei

3 Die Auswirkung alternativer Inventarumfänge (also ohne Umlaute bzw. ohne ß) wurde noch nicht im Detail untersucht; die entsprechenden Resultate werden an anderer Stelle darzustellen sein.

Häufigkeitsverteilungen sind es Funktionen wie Variationskoeffizient, Schiefe, oder Exzess (die in der Linguistik eher selten benutzt werden), die Momente von Verteilungen oder auch Eigenschaften wie (die approximativ ineinander überführbare) Wiederholungsrate und Entropie und andere mehr, die man recht häufig antrifft. Die Berechnung solcher Eigenschaften eignet sich also zur Beantwortung der Frage, ob es auf der globalen Ebene zu Einwirkungen kommt.
Am bekanntesten ist vermutlich die Entropie, die im Rahmen der Informationstheorie seit Shannon in der Regel als ein spezifisches Informationsmaß definiert worden ist, und zwar nach der Formel

(1) $$H_1 = -\sum_{i=1}^{I} p_i \cdot \mathrm{ld}\ p_i \ .$$

Die Entropie lässt sich als ein Maß der Gleichverteilung verstehen: je größer die Entropie, desto einander ähnlicher die Vorkommenshäufigkeiten p_i. Es lässt sich leicht zeigen, dass H_1 sich im Intervall <0; ld I> bewegt, d.h. das theoretische Minimum der Entropie beträgt $H_1 = 0$, und das Maximum ist $H_1 = \mathrm{ld}\ I$, wobei I hier den Inventarumfang bezeichnet.
Damit hängt H_1 ganz klar vom Inventarumfang I ab. Dieser Umstand wird oft nicht angemessen berücksichtigt, wenn die Entropien zweier Stichproben miteinander verglichen werden. Wenn wir also die Entropiewerte unserer Stichproben für die verschiedenen Bedingungen miteinander vergleichen wollen, dann ist von vornherein zu erwarten, dass die Entropie für D-30 größer sein müsste als für D-29 – was den Vergleich überflüssig erscheinen ließe. Insofern liegt es nahe, in Kenntnis des theoretischen Minimums und des theoretischen Maximums H_1 zu relativieren und zum Zwecke des Vergleichs die relative Entropie zu berechnen:

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

Dies hat den weiteren Vorteil, dass sich nun auf der Basis von Formel auch die Redundanz der jeweiligen Stichproben nach der folgenden Formel berechnen und miteinander vergleichen lässt:

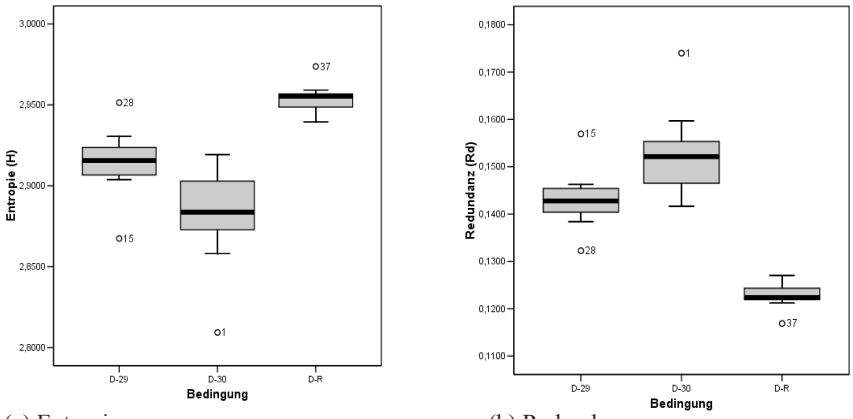
(3) $$Rd = 1 - H_{\mathrm{rel}} \ .$$

Tab. 3 enthält die entsprechenden Werte für die drei Bedingungen.

Tab. 3: Entropie und Redundanz für D-30, D-29, und D-R

	D-30				D-29				D-R		
	H	*H*rel	*Rd*		*H*	*H*rel	*Rd*		*H*	*H*rel	*Rd*
1	2,8094	0,8260	0,1740	**1**	2,8674	0,8431	0,1569	**1**	2,9498	0,8760	0,1240
2	2,8729	0,8447	0,1553	**2**	2,9162	0,8574	0,1426	**2**	2,9396	0,8730	0,1270
3	2,9154	0,8572	0,1428	**3**	2,9146	0,8569	0,1431	**3**	2,9486	0,8757	0,1243
4	2,8654	0,8425	0,1575	**4**	2,9037	0,8537	0,1463	**4**	2,9568	0,8781	0,1219
5	2,8582	0,8403	0,1597	**5**	2,9067	0,8546	0,1454	**5**	2,9592	0,8788	0,1212
6	2,8744	0,8451	0,1549	**6**	2,9070	0,8547	0,1453	**6**	2,9566	0,8780	0,1220
7	2,8854	0,8484	0,1516	**7**	2,9280	0,8609	0,1391	**7**	2,9560	0,8779	0,1221
8	2,8845	0,8481	0,1519	8	2,9237	0,8596	0,1404	**8**	2,9547	0,8775	0,1225
9	2,8853	0,8483	0,1517	**9**	2,9057	0,8543	0,1457	**9**	2,9737	0,8831	0,1169
10	2,9046	0,8540	0,1460	**10**	2,9204	0,8586	0,1414	**10**	2,9426	0,8739	0,1261
11	2,8829	0,8476	0,1524	**11**	2,9157	0,8573	0,1427				
12	2,9029	0,8535	0,1465	**12**	2,9305	0,8616	0,1384				
13	2,8813	0,8471	0,1529	**13**	2,9154	0,8572	0,1428				
14	2,9194	0,8583	0,1417	**14**	2,9514	0,8677	0,1323				

Abb. 1 veranschaulicht gegenüberstellend das Verhalten der absoluten Entropie und der (sich aus der relativen Entropie ergebenden) Redundanz, die im gegebenen Fall zu ein und derselben Aussage führen: Demnach ist zunächst einmal – entsprechend unserer Erwartung – die Entropie für D-30 kleiner als für D-29. Andererseits aber ist die Entropie unter der Bedingung D-R noch größer als unter den beiden anderen Bedingungen, d.h. auch deutlich größer als unter der Bedingung D-29, was insofern besonders interessant ist, als diesen beiden Bedingungen ja ein identischer Inventarumfang von $I = 29$ zugrunde liegt.



(a) Entropie (b) Redundanz

Abb. 1: Entropie und Redundanz unter den drei Bedingungen (D-29, D-30, D-R)

Der Test der Unterschiede auf statistische Signifikanz in Form des nicht-parametrischen Kruskall-Wallis-H-Tests weist die Unterschiede sowohl für die absolute Entropie $X^2 = 26.66$ als auch für die Redundanz ($X^2 = 27.54$) bei jeweils zwei Freiheitsgraden als hoch signifikant aus ($p < 0.001$); auch die paarweisen Vergleiche (D-30 vs. D-29, D-30 vs. D-R, sowie D-29 vs. D-R) in Form von Mann-Whitney-U-Tests führen zu demselben Ergebnis.
Mit diesem Befund ist somit der Nachweis erbracht, dass sich das System der Buchstabenhäufigkeiten durch die Einführung der Restriktion nachhaltig ändert: Die Redundanz des gesamten Systems wird wesentlich reduziert, was sich so interpretieren lässt, dass sich aufgrund der stärkeren Gleichverteilung der Buchstaben eine höhere funktionale Auslastung der verbliebenen Systemelemente ergibt. Dieser Befund lässt deshalb auch Unterschiede in der Art der Häufigkeitsverteilung der Buchstaben erwarten, die es im weiteren detailliert zu untersuchen gilt.

Ein Modell der Organisation von Graphemhäufigkeiten: Die negative hypergeometrische Verteilung

Die Frage nach der Organisation der Häufigkeitsverteilung ist darauf ausgerichtet zu untersuchen, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, diese 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, wobei man wie folgt vorgeht: Überführt man erhobene Ausgangsdaten in eine Rang-Reihenfolge, sortiert die Häufigkeiten in absteigender Reihenfolge und verbindet sodann die Datenpunkte miteinander, dann ergibt sich charakteristischerweise kein linearer Abfall, sondern eine spezifische, monoton fallende (üblicherweise hyperbolische) Kurve. Und genau darum geht es nun im Prinzip: nämlich die genaue Form dieser Kurve zu modellieren, um so zu sehen, ob die Häufigkeiten (d.h. die spezifische Abnahme der Häufigkeiten) in verschiedenen Stichproben ein und dieselbe Form aufweisen oder nicht. Allerdings haben wir es bei unseren Daten bei näherem betrachtet nicht mit einer stetigen (kontinuierlichen) Kurve, sondern mit einem diskreten System tun, und damit mit der Differenz von zwei jeweils benachbarten Größen. An diesem Punkt liegt es nahe, auf Grundannahmen der synergetischen Linguistik zu rekurrieren.

Hier wird bei Verteilungsproblemen im allgemeinen angenommen, dass die Vorkommenswahrscheinlichkeit einer bestimmten Klasse P_x mit der Ausprägung x oder dem Rang r sich proportional zu der jeweils niedrigeren Klasse (also x–1 bzw. r–1) entwickelt (vgl. Altmann/Köhler 1996). Von diesem allgemeinen Ansatz ausgehend kann man also die Differenzengleichung

$$P_x = g(x)P_{x-1} \tag{4}$$

aufstellen, deren konkrete Lösung von der jeweiligen Funktion *g(x)* abhängt. Bereits einfache Funktionen haben in der Vergangenheit bei der Untersuchung der Vorkommenshäufigkeit verschiedener sprachlicher Einheiten zu überzeugenden Resultaten geführt. Dabei war *g(x)* üblicherweise eine einfache gebrochene rationale Funktion, bei deren qualitativer Interpretation man davon ausging, dass sie im Zähler die „Kräfte" des Sprechers, im Nenner die regulierenden „Kräfte" des Hörers enthielt. Wimmer/Altmann (2005, 2006) verallgemeinerten diesen Ansatz nunmehr in zwei Richtungen:

(a) Im ersten Fall gibt es nur eine einzige unabhängige Variable, nämlich x; alle anderen werden als ceteris paribus betrachtet; da sie aber nicht alle gleich sind und keine Konstante darstellen, werden sie gewichtet, und zwar als b/x^2, c/x^3, ...

(b) Im zweiten Fall gibt es mehrere unabhängige Variablen, die alle explizit berücksichtigt werden, wodurch sich partielle Differenzen- und Differentialgleichungen ergeben.

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

$$P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2}\right)P_{x-1} \tag{5}$$

$$P_x = \left(1 + a_0 + \frac{a_1}{x + b_1} + \frac{a_2}{(x + b_2)^2}\right)P_{x-1} \tag{6}$$

$$P_x = \left(1 + a_0 + \frac{a_1}{(x + b_1)^{c_1}} + \frac{a_2}{(x + b_2)^{c_2}}\right)P_{x-1} \tag{7}$$

Die Lösungen dieser Gleichungen sind vergleichsweise einfach: Man sieht sofort, dass (5) ein Spezialfall von (6), und (6) ein Spezialfall von (7) ist. Aus diesem Ansatz lassen sich u.a. verschiedene Verteilungsmodelle ableiten. Ein Modell, dessen Eignung zur Modellierung von Buchstabenhäufigkeiten in

jüngster Zeit mehrfach empirisch nachgewiesen wurde, ist die negative hypergeometrische Verteilung (Wimmer/Altmann 1999: 465-468), die auch unter der Bezeichnung Beta-Binomial-Verteilung bekannt ist. Man kann diese Verteilung auf verschiedene Weisen ableiten – eine Herleitung aus Ansatz stellt sich wie folgt dar: Setzt man hier

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

so erhält man die Rekursionsformel

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

woraus die negative hypergeometrische Verteilung resultiert:

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

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

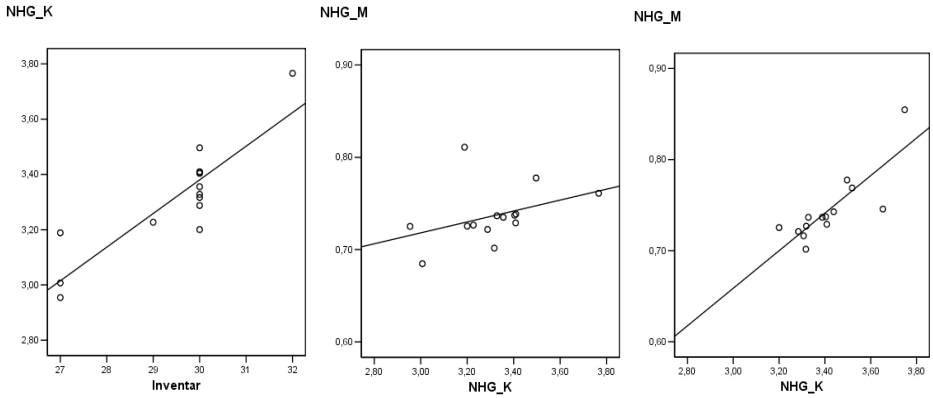
(10) $$P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}} \qquad \begin{aligned} &x = 1,2,\ldots,\ n+1\\ &K > M > 0;\ n \in \{1,2,\ldots\}\end{aligned}$$

Die Eignung dieses Verteilungsmodells für Buchstabenhäufigkeiten ist in jüngster Zeit in den oben erwähnten Untersuchungen an slawischen Sprachen wiederholt nachgewiesen worden; aufgrund von systematischen Untersuchungen konnten zudem erste Einsichten in die Systematik des Verhaltens der Parameter *K* und

M gewonnen werden; einerseits scheint demnach übersprachlich eine (lineare) Abhängigkeit des Parameters *K* vom Inventarumfang *I* vorzuliegen, andererseits innerhalb der jeweils einzelnen Sprachen eine lineare Abhängigkeit des Parameter *M* von *K*. Diese Verhältnisse lassen sich für die bislang untersuchten slawischen Sprachen (Russisch, Slowakisch, Slowenisch, Ukrainisch) auf ein ihnen gemeinsam zugrunde liegendes Regressionsmodell zurückführen, innerhalb dessen sich sprachübergreifende ebenso wie sprachspezifische Prinzipien differenziert betrachten lassen (Grzybek/Kelih/Stadlober 2006). Versuche einer weiterführenden inhaltlichen Interpretation der Parameter stehen erst am Anfang, doch liegen erste Hinweise im Hinblick auf eine besondere Bedeutung des Mittelwerts einer gegebenen Verteilung für die Parameterwerte vor.

So weit ist die Untersuchung anderer Sprachen bislang noch nicht vorgedrungen. Das gilt auch für die erwähnte Untersuchung von Best (2005) zum Deutschen, in der aufgrund der Analyse von 14 Stichproben auch für das Deutsche die negative hypergeometrische Verteilung als geeignetes Modell ins Spiel gebracht wurde: in der Tat liegt der sich als X^2/N berechnende Diskrepanzkoeffizient *C* – der bei $C < 0.02$ als Ergebnis einer guten, bei $C < 0.01$ einer sehr guten Übereinstimmung zwischen Daten und Modell angesehen wird –, nach Ausschluss einer mit nur 675 Buchstaben extrem kleinen Stichprobe bei einem Mittelwert von $\bar{x}_c = 0.0079$ im Intervall von $0.0046 \leq C \leq 0.0128$. Allerdings sind die Ergebnisse bislang weit von einer Interpretation entfernt, zumal die von Best (2005) für *K* und *M* erhaltenen Parameterwerte keine Systematik aufweisen. Grund dafür dürfte der Umstand sein, dass der Untersuchung der einzelnen Stichproben jeweils unterschiedliche Inventarumfänge zugrunde liegen: Damit variiert der Parameter *n* der NHG von 27 (wenn etwa in einer Stichprobe die Buchstaben Q, X, Y nicht vorkommen), über 29 (wenn der Buchstabe ß nicht vorkommt), bis hin zu 30, in einem Fall sogar bis zu 32, da in diesem Text (*Winnetou I* von Karl May) zufällig É und Ñ vorkommen und entsprechend als Inventarbestandteile berücksichtigt wurden).

Interessanterweise konnte Grzybek (2007) allerdings nachweisen, dass sich sogar unter diesen Voraussetzungen, trotz der geringen Anzahl an Stichproben innerhalb ein und derselben Sprache recht deutlich die an den slawischen Sprachen wiederholt beobachtete Abhängigkeit des Parameters *K* vom Inventarumfang *I*, wie Abb. 2a zeigt.



(a) *K* und *I* (b) *K* und *M* (nach Best 2005) (c) *K* und *M* (D-30; nach Grzybek 2007)

Abb. 2: Parameter *K*: Abhängigkeiten und Zusammenhänge

Allerdings zeigt Abb. 2b ebenso deutlich, wie es sich auswirkt, dass der Inventarumfang nicht auf der Basis des jeweiligen Systemumfangs definiert, sondern den in den konkreten Stichproben jeweils beobachteten Umfang angeglichen wurde: Wie zu sehen ist, gibt es unter dieser Bedingung keinen erkennbaren Zusammenhang zwischen den Parametern *K* und *M* . Es liegt auf der Hand, dass sich die Schwankung von *I* (in der NHG dem Parameter *n* entsprechend) auch auf das Verhältnis der Parameterwerte von *K* und *M* auswirkt, das sich somit einer Interpretation entzieht. Legt man hingegen den Inventarumfang bei $I = 30$ fest[4], so zeigt sich eine klare Korrelation zwischen den Parametern *M* und *K* ($r = 0.921$, $p < 0.001$) – vgl. Abb. 2c.

Damit ist erstmal der Nachweis eines systematischen Parameterverhaltens auch für deutsche Buchstabenhäufigkeiten erbracht. Im hier gegebenen Kontext muss darauf verzichtet werden, dieses zu den an den bislang untersuchten slawischen Sprachen gewonnenen Ergebnissen in Bezug zu setzen. Statt dessen bietet es sich an, die Eignung dieses Modells auch an den sich aus der deutschen Perec-

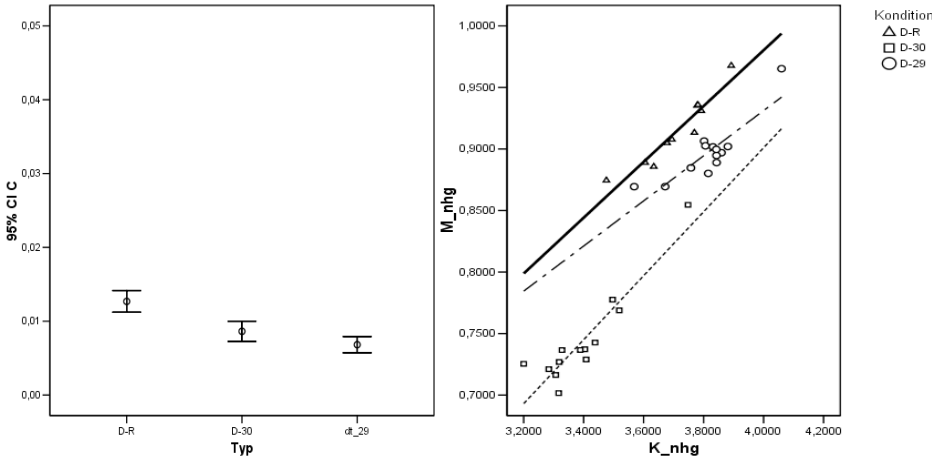
4 Dieser Definition gemäß beinhaltet das System der deutschen Buchstaben sowohl die Umlaute Ä,Ö,Ü als auch das ß als eigenständige Elemente. In einer detaillieren Re-Analyse zur Auswirkung der festgelegten Inventargröße wurden die Daten von Best (2005) alternativ auch mit $I = 26$ und $I = 27$ ausgewertet (Grzybek 2007); dabei wurden systematisch die Umlaute in Doppelvokale (Ä→AE, Ö→OE, Ü→UE) gewandelt, die optionale Wandlung von ß→SS wurde als weitere Variante behandelt. Dieses Vorgehen erlaubt neben einem Vergleich der Anpassungsgüte auch wiederum einen Untersuchung des Zusammenhangs von *K* und *I* innerhalb des Deutschen unter verschiedenen Bedingungen; ein detaillierteres Eingehen auf die Auswirkungen würde jedoch weit über den hier gegebenen Rahmen hinausgehen.

Übersetzung ergebenden Daten (D-R) zu testen, die Ergebnisse auf allfällige Modifikationen im Vergleich zu den Bedingungen D-30 und D-29 hin zu prüfen und gegebenenfalls zu interpretieren. Tab. 4 enthält die entsprechenden Parameterwerte für *K* und *M* sowie die *C*-Werte der Anpassungsergebnisse.

Tab. 4: Anpassungsergebnisse (NHG – D-30, D-29, D-R)

D-30				D-29				D-R			
	K	*M*	*C*		*K*	*M*	*C*		*K*	*M*	*C*
1	3,3071	0,7163	0,0110	1	3,7576	0,8846	0,0077	**1**	3,6943	0,9077	0,0144
2	3,7480	0,8546	0,0125	**2**	4,0597	0,9652	0,0107	**2**	3,7691	0,9132	0,0108
3	3,4083	0,7289	0,0100	**3**	3,8431	0,8889	0,0075	**3**	3,7922	0,9310	0,0152
4	3,3167	0,7016	0,0073	**4**	3,8154	0,8801	0,0051	**4**	3,7780	0,9353	0,0112
5	3,4381	0,7427	0,0102	**5**	3,8595	0,8970	0,0080	**5**	3,8919	0,9677	0,0140
6	3,2839	0,7210	0,0100	**6**	3,8015	0,9064	0,0060	**6**	3,7810	0,9358	0,0112
7	3,3184	0,7269	0,0104	**7**	3,8059	0,9026	0,0073	**7**	3,6778	0,9049	0,0107
8	3,5189	0,7688	0,0092	**8**	3,8807	0,9020	0,0082	**8**	3,6050	0,8888	0,0118
9	3,4964	0,7775	0,0066	**9**	3,8309	0,9017	0,0062	**9**	3,4758	0,8744	0,0114
10	3,4038	0,7372	0,0058	**10**	3,8428	0,8947	0,0050	**10**	3,6336	0,8856	0,0162
11	3,3278	0,7366	0,0052	**11**	3,6717	0,8694	0,0056				
12	3,3886	0,7366	0,0062	**12**	3,8423	0,8996	0,0036				
13	3,2002	0,7254	0,0079	**13**	3,5684	0,8694	0,0077				

Abb. 3a veranschaulicht die Tendenz der Ergebnisse: wie zu sehen ist, stellt die NHG auch für die Restriktionsbedingung (D-R) ein geeignetes Modell dar: Die *C*-Werte liegen bei einem Durchschnitt von $\bar{x}_c = 0.0127$ im Intervall von $0.0107 \leq C \leq 0.0162$; die Anpassungsgüte ist dabei insgesamt nur unwesentlich schlechter als für die beiden Bedingungen D-29 und D-30.
Besonders aufschlussreich ist nun ein neuerlicher Vergleich des Parameterverhaltens unter den drei Bedingungen, das in Abb. 3b anschaulich dargestellt ist; hervorzuheben sind



(a) *C*-Werte (b) Abhängigkeit von *K* und *M*

Abb. 3: Anpassungsergebnisse (NHG – D-30, D-29, D-R)

1. die durchgehend lineare Abhängigkeit des Parameters *M* von *K* innerhalb der jeweiligen Bedingung,
2. der auch schon an den bislang untersuchten slawischen Sprachen beobachtete Parallelitätstendenz der Regressionsgeraden,
3. der Unterschied der Ausprägung des Wertes für *K* unter der Restriktionsbedingung im Vergleich zu D-29 trotz identischem Inventarumfang,
4. die geringere Variabilität und noch stärker ausgeprägte Korrelation der Beziehung zwischen den Parametern *K* und *M* für D-R,
5. der relativ hohe Parameterwert von *M* unter der Bedingung D-R, der eine klare Tendenz aufweist, sich dem Wert $M = 1$ zu nähern.

Insbesondere die letzte Beobachtung gibt Anlass zu der Vermutung, dass für die Bedingung D-R auch ein einfacheres Modell als die NHG bzw. ein Spezialfall selbiger zur Modellierung der Häufigkeiten ausreichen könnte. Sollte dies der Fall sein, würde die Aufgabe der Parameter-Interpretation natürlich wesentlich erleichtert, da nur ein Parameter zur Interpretation übrig bliebe, denn im Falle von $M = 1$ ergibt sich anstelle der in dargestellten NHG die sogenannte Dacey-3-Verteilung:

$$P_x = \frac{\binom{m+n-x-2}{n-x}}{\binom{K+n-1}{n}} \quad . \qquad x = 0,1,2,\ldots,n, \quad m \in \mathbb{N},\ n \in \mathbb{N}_0 \tag{11}$$

bzw., in 1-verschobener Form:

(12) $$P_x = \frac{\binom{m+n-x-1}{n-x+1}}{\binom{m+n-1}{n}} \quad . \qquad x = 1,2,\ldots,n+1$$

Auch diese Verteilung lässt sich natürlich aus der oben angesprochenen »Unified Theory« von Wimmer/Altmann (2005, 2006) ableiten, wobei in diesem Fall sogar deren erste drei Glieder ausreichen:

(13) $$P_x = \left(1 + a_0 + \frac{a_1}{x-b}\right) P_x$$

Ersetzt man hier $a_0 = 0$, $a_1 = m - 2$, und $b = m + n - 1$, erhält man die Rekursionsformel der Dacey-3-Verteilung:

(14) $$P_x = \frac{n-x+1}{m+n-x-1} P_{x-1} \quad .$$

Dieses Verteilungsmodell ist in der quantitativen Linguistik bislang noch nie angewendet worden. Tab. 5 enthält die Ergebnisse der Anpassung im Vergleich mit D-29 und D-30.

Tab. 5: Anpassungsergebnisse der Dacey-3-Verteilung (D-30, D-29, D-R)

	D-30			D-29			D-R	
	K	C		*K*	C		*K*	C
1	4,7253	0,0568	**1**	4,3778	0,0135	**1**	4,1988	0,0163
2	4,5233	0,0225	**2**	4,3562	0,0125	**2**	4,2716	0,0136
3	4,7728	0,0505	**3**	4,4469	0,0123	**3**	4,2062	0,0168
4	4,7800	0,0588	**4**	4,4234	0,0101	**4**	4,1774	0,0128
5	4,7211	0,0459	**5**	4,4204	0,0120	**5**	4,1988	0,0138
6	4,6494	0,0544	**6**	4,3077	0,0095	**6**	4,1787	0,0128
7	4,6528	0,0523	**7**	4,3272	0,0110	**7**	4,1792	0,0132
8	4,6642	0,0364	**8**	4,4145	0,0120	**8**	4,1860	0,0168
9	4,5713	0,0312	**9**	4,3493	0,0095	**9**	4,0888	0,0169
10	4,6871	0,0429	**10**	4,3982	0,0082	**10**	4,2354	0,0213
11	4,5745	0,0422	**11**	4,3159	0,0114			
12	4,6990	0,0438	**12**	4,3894	0,0078			
13	4,4689	0,0499	**13**	4,1943	0,0137			

In der die Ergebnisse veranschaulichenden Abb. 4 sind zum Zwecke des Vergleichs ebenfalls die Werte für die NHG eingezeichnet (s.o.)

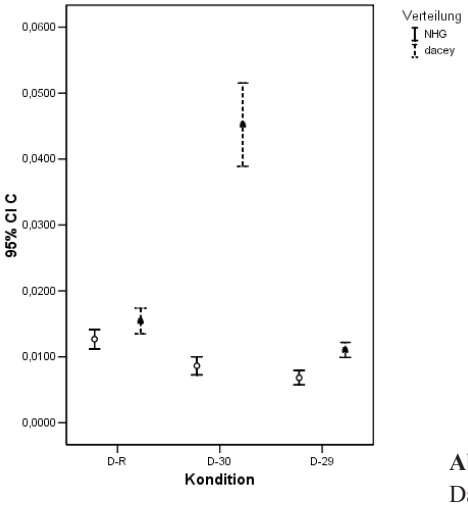


Abb. 4: Anpassungsergebnisse (NHG vs. Dacey-3)

Deutlich zeigt sich, dass die Dacey-3-Verteilung für D-R ebenso wie für D-29 ein überaus geeignetes Modell ist, das im Vergleich zur NHG nur unwesentlich schlechter ist, aufgrund der geringeren Parameteranzahl aber vorzuziehen wäre. Ebenso deutlich zeigt sich, dass diese Verteilung für die „normalen" Häufigkeiten (D-30) vollkommen ungeeignet ist. Vor diesem Hintergrund sind die wechselseitigen Vergleiche zwischen den drei Bedingungen und der Eignung der beiden Verteilungsmodelle geradezu richtungweisend, insofern zwei verschiedene Beobachtungen zu ein und derselben Schlussfolgerung führen; wenn nämlich die Dacey-3-Verteilung

- sowohl für D-R als auch (sogar noch besser) für D-29 geeignet ist, und
- gleichzeitig für die „normale" Bedingung D-30 ungeeignet ist, letztere aber durch die NHG gut erfasst werden kann,

dann spricht dies in der Zusammenschau dieser Befunde dafür, dass es gerade die erste Häufigkeit einer „normalen" Häufigkeitsverteilung deutscher Buchstaben ist, welche durch die NHG (bzw. deren zusätzlichen Parameter) erfasst wird. Diese Schlussfolgerung gibt in weiterer Folge Anlass zu einer Neu-Interpretation der NHG, d.h. der Schätzung und inhaltlichen Interpretation ihrer Parameter. Diesem Hinweis wird an anderer Stelle im Detail nachzugehen sein. Im hier gegebenen Kontext bietet sich allerdings noch eine andere, überaus at-

traktive Variante an: Diese besteht in der Entwicklung und Überprüfung einer Modifikation der Dacey-3-Verteilung, wie sie sich in Formel darstellt:

$$(15) \qquad P_x = \begin{cases} 1-\alpha, & x=1 \\ \dfrac{\alpha\binom{m+n-x}{n-x+2}}{\binom{m+n-1}{n}}, & x=2,3,\ldots,n+2 \end{cases}$$

Hier wird die gesamte Verteilung nicht, wie in , um nur einen, sondern um zwei Schritte zunächst nach rechts verschoben und dann mit α multipliziert, wobei die erste Häufigkeit (P_1) als $P_1 = 1 - \alpha$ festgelegt wird (und folglich $\alpha = 1 - P_1$). Im Prinzip weist also die modifizierte Dacey-3-Verteilung ebenso wie die NHG neben dem Parameter n noch zwei weitere auf; allerdings hätte die Dacey-3-Verteilung den Vorteil, dass einer von diesen (α) durch die Bezugnahme auf P_1 eine Erklärung aufweisen würde und somit nur ein weiterer einer Interpretation bedürfte.[5] Ein weiterer Vorteil der Dacey-3-Vorteilung ist, dass sich ihr Parameter m sehr leicht über den empirischen Mittelwert m_1 der gegebenen Verteilung schätzen lässt, da der theoretische Mittelwert $\hat{m}_1$ der Dacey-3-Verteilung gegeben ist als

$$(16) \quad \hat{m}_1 = \frac{m}{n} + 1 \;;$$

dies führt folglich zu der Schätzung von m als

$$(17) \qquad m = \frac{n}{m_1 - 1} \quad .$$

Tab. 6 enthält die Parameterwerte für m, und zwar sowohl für die iterativ nach der Methode der Chi-Quadrat-Minimierung erhaltenen Werte (m_{iter}) als auch die nach Formel (17) erhaltenen Schätzwerte (m_{est}).

5 Natürlich würde das Vorgehen insgesamt einer inhaltlichen Begründung bedürften; diese müsste der Tatsache Rechnung tragen, dass gerade die erste Häufigkeit vom allgemeinen Verhalten aller übrigen abweicht und deswegen einer statistischen „Spezialbehandlung“ bedarf – Gründe dafür könnte eine funktionale Mehrauslastung des häufigsten Buchstabens sein (etwa der Repräsentation mehrerer Phoneme zu dienen), die Tatsache, dass sich die für das Funktionieren eines Systems notwendige Entropie primär über das häufigste Element erreichen lässt, u.a.m.– Freilich muss eine inhaltliche Begründung zukünftigen Untersuchungen vorbehalten bleiben.

Tab. 6: Anpassungsergebnisse (mod. Dacey – D-30, D-29, D-R)

D-30				D-29				D-R			
	m_{est}	m_{iter}	C		m_{est}	m_{iter}	C		m_{est}	m_{iter}	C
1	4,7760	4,5319	0,0112	**1**	4,3994	4,5319	0,0135	**1**	4,2153	4,2701	0,0122
2	4,5369	4,5011	0,0119	**2**	4,3562	4,5011	0,0138	**2**	4,2746	4,3559	0,0112
3	4,8372	4,6064	0,0100	**3**	4,4778	4,6064	0,0121	**3**	4,2189	4,2695	0,0133
4	4,8906	4,5834	0,0077	**4**	4,4898	4,5834	0,0094	**4**	4,1837	4,2632	0,0109
5	4,7891	4,5794	0,0099	**5**	4,4565	4,5794	0,0118	**5**	4,1654	4,2612	0,0134
6	4,7116	4,4590	0,0080	**6**	4,3438	4,4590	0,0096	**6**	4,1845	4,2640	0,0109
7	4,7227	4,4800	0,0091	**7**	4,3673	4,4800	0,0109	**7**	4,2095	4,2474	0,0084
8	4,7347	4,5704	0,0100	**8**	4,4560	4,5704	0,0119	**8**	4,2007	4,2133	0,0091
9	4,6518	4,5048	0,0078	**9**	4,4004	4,5048	0,0092	**9**	4,1168	4,1168	0,0073
10	4,7764	4,5591	0,0062	**10**	4,4485	4,5591	0,0075	**10**	4,2494	4,2537	0,0121
11	4,6736	4,4754	0,0088	**11**	4,3741	4,4754	0,0104				
12	4,7590	4,5420	0,0066	**12**	4,4235	4,5420	0,0079				
13	4,5639	4,3475	0,0109	**13**	4,2513	4,3475	0,0130				

Wie zu sehen ist, weichen die geschätzten und iterativ bestimmten *m*-Werte durchgehend nur geringfügig voneinander ab. Auch die jeweiligen *C*-Werte unterscheiden sich entsprechend nur geringfügig, angeführt ist der nach Iteration erhaltene. Deutlich zu sehen ist, dass die modifizierte Dacey-3-Verteilung für alle drei Bedingungen ein bestens geeignetes Modell darstellt (vgl. Abb. 5).

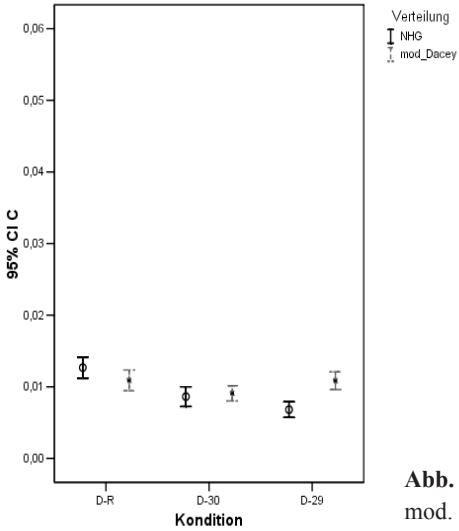


Abb. 5: Anpassungsergebnisse (NHG vs. mod. Dacey-3)

Besonders überraschend ist im Vergleich der Befund, dass sich für die „normale“ Bedingung D-30 die Dacey-3-Verteilung als ebenso geeignet erweist wie die NHG. Diese Beobachtung lässt zumindest zwei Schlussfolgerungen zu:

1. Ab sofort steht mit der Dacey-3-Verteilung zumindest für deutsche Buchstabenhäufigkeiten ein Modell zur Verfügung, das als Spezialfall der NHG im Vergleich zu dieser dadurch an Charme gewinnt, dass hier nicht nur der erste Parameter (α) eine Erklärung findet, sondern auch der zweite (m), insofern dieser unmittelbar auf den Mittelwert der Verteilung zurückzuführen ist.
2. Die Beobachtung über die Bedeutung des Mittelwertes – deren Bedeutung ansatzweise schon in den Untersuchungen zu slawischen Sprachen vermutet wurde – öffnet die Möglichkeit einer Neu-Interpretation der Parameter der NHG, wobei sich aufgrund der in dieser Darstellung vorgetragenen Beobachtungen ein Zusammenspiel zwischen der ersten Häufigkeit einer Verteilung (P_1), ihrem Mittelwert (m_1) und dem jeweiligen Inventarumfang (n) ergibt, was schließlich die Eignung der NHG und die Notwendigkeit eines derartig komplexen Modells erklären könnten.

Resümee

In Vordergrund der vorgelegten Untersuchung steht die Analyse der Vorkommenshäufigkeit von Buchstaben und damit einhergehender Prozesse der Selbst-Regulation. Ausgehend von der Kenntnis um Prinzipien der Häufigkeitsorganisation wird an spezifischem Textmaterial, nämlich dem »Roman ohne E« von George Perec, die Auswirkung der eingeführten Restriktion auf die Organisation der Häufigkeitsstruktur analysiert.

Es bestätigt sich zunächst die Annahme, dass sich das System der Buchstabenhäufigkeiten durch die Einführung der Restriktion nachhaltig ändert: Die Redundanz des gesamten Systems wird wesentlich reduziert, einhergehend mit einer höheren funktionalen Auslastung der verbliebenen Systemelemente.

Die Einführung der Restriktion führt dabei keineswegs zu irgendeiner Art von Chaosbildung, sondern mündet in einen Prozess der Selbst-Organisation, der im Prinzip auch unter „normalen“ Bedingungen stattfindet. Entsprechend lassen sich die Häufigkeiten deutscher Buchstaben unter beiden Bedingungen theoretisch durch ein und dasselbe Modell, die negative hypergeometrische Verteilung (NHG) beschreiben, das sich auch schon an anderen (slawischen) Sprachen wiederholt als geeignet erwiesen hat;

Interessanterweise gibt allerdings gerade die experimentelle Analyse der Auswirkung der eingeführten Restriktion Hinweise auf ein einfacheres Modell, das einen Spezialfall der ansonsten sehr komplexen NHG darstellt. Dieses Modell – die sog. Dacey-3-Verteilung – eignet sich allerdings offenbar nicht für die Modellierung unter „normalen“ Voraussetzungen (D-30), sondern nur für die beiden Experimentalbedingungen (d.h. D-29 und D-R). Mit diesem Befund ist ein eindeutiger Hinweis auf die besondere Bedeutung der ersten Häufigkeit (P_1) einer Ranghäufigkeit gegeben, die entweder nur durch das komplexere Modell der NHG erfasst werden kann, oder aber durch eine Modifizierung der Dacey-3-Verteilung, die eine „Spezialbehandlung“ eben der ersten Häufigkeit vorsieht.
In der Zusammenschau der Ergebnisse zeichnet sich somit ein Bild ab, dem zufolge in Modelle für die theoretische Beschreibung von Buchstabenhäufigkeiten offenbar auf die eine oder andere Art und Weise drei Größen integriert sein müssen: die erste Häufigkeit (P_1), der Mittelwert (m_1) und der jeweiligen Inventarumfang (I). Vor diesem Hintergrund wird die hier vorgestellte modifizierte Dacey-3-Verteilung in Hinkunft bei weiteren Analysen zu Buchstabenhäufigkeiten – zumindest im Deutschen – unbedingt als eine Option zu berücksichtigen sein.
Es scheint jedenfalls, dass der Code der Buchstabenhäufigkeiten nicht unknackbar ist, und wenn mit der vorliegenden Untersuchung ein wesentlicher Schritt in diese Richtung getan ist, hat die Sprachpolitik des Oulipo eines ihrer Ziele auf einem nicht gerade kurzen Umweg erreicht, nämlich die Einsicht in Prozesse der sprachlichen Organisation durch die Erforschung der Auswirkung zusätzlicher eingeführter Restriktionen.

Primärliteratur

Perec, Georges (1969): *La disparition*. Paris.

Übersetzungen
Deutsch:
*Anton Voyls Fortgang***.** Hrsg. u. übers. von Eugen Helmlé. Frankfurt/M. 1986.

Englisch:
A Void. Trans. Gilbert Adair. London 1994.

Russisch:
Isčezanie. Perevod Valerij Kislov. Sankt-Peterburg 2005.

Sekundärliteratur

Chesneau, A. (1984): „Lorsque l'E disparaît du cercle des voyelles", in: *Francofonia: studi e ricerce sulle letterature di lingua francese* (4), 5-16.

Helmlé, E. (1977): „Anton Voyl schwand dahin. Über das Leipogramm und Georges Perecs *La Disparition*", in: *Akzente* (24/5), 453-472.

Kislov, V. (2006): „Traduire *La disparition* de Perec en russe", in: *Formules. Revue des littératures à contraintes* (10), 279-308.

Perec, G. (1973): „Histoire du lipogramme." In: Oulipo: *La littérature potentielle.* Paris, 77-93.

Schleypen, U. (2004): *Schreiben aus dem Nichts: Gegenwartsliteratur und Mathematik. Das Ouvroir de littérature potentielle.* München.

Altmann, G.; Levickij, V.; Perebejnis, V. (Hrsg.) (2005): *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics.* Černovici.

Best, K.-H. (2005): „Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten", in: *Glottometrics* (11), 9-31.

Galland, J.S. (1941): „»E«. A Study of the Structural Elements of the French Language", in: *The Modern Language Journal* (25/6), 477-487.

Garabík, R. (Hrsg.) (2005): *Computer Treatment of Slavic and East European Languages.* Bratislava.

Grzybek, P. (2005): „A Study on Russian Graphemes." In: Toporov (Hrsg.) (2005), 237-263.

Grzybek, P. (Hrsg.) (2006): *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht.

Grzybek, P. (2007b): „On the Systematic and System-based Study of Grapheme Frequencies (A Re-Analysis of German Letter Frequencies)", in : Glottometrics [in Druck].

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

Grzybek, P.; Kelih, E. (2005a): „Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph." In: Altmann/Levickij/Perebejnis (Hrsg.) (2005), 159-179.

Grzybek, P.; Kelih, E. (2005b): „Towards a General Model of Grapheme Frequencies in Slavic Languages. In: Garabík (Hrsg.) (2005), 73-87.

Grzybek, P.; Kelih, E.; Altmann, G. (2004): „Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung", in: *Anzeiger für Slavische Philologie* (XXXII), 25-54.

Grzybek, P.; Kelih, E.; Altmann, G. (2005a): „Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das 'ё'", in: *Anzeiger für Slavische Philologie* (XXXIII), 117-140.

Grzybek, P.; Kelih, E.; Altmann, G. (2005b): „Graphemhäufigkeiten im Slowakischen. (Teil I: Ohne Digraphen)." In: Nemcová (Hrsg.)

Grzybek, P.; Kelih, E.; Altmann, G. (2006a): „Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen." In: Kozmová (Hrsg.) (2006), 661-664.

Grzybek, P.; Kelih; E.; Stadlober, E. (2006b): „Graphemhäufigkeiten des Slowenischen (und anderer

slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik“, in: *Anzeiger für Slavische Philologie* (XXXIV), 41-74.

Kozmová, R. (Hrsg.) (2006): *Sprache und Sprachen im mitteleuropäischen Raum.* Trnava.

Köhler, R.; Altmann, G.; Piotrowski, R. G. (Hrsg.) (2005): *Quantitative Linguistics. An International Handbook.* Berlin, New York.

Nemcová, E. (Hrsg.), *Philologia actualis slovaca*. [in Druck]

Schönpflug, W. (1969): „*n*-Gramm Häufigkeiten in der deutschen Sprache. I: Monogramme und Digramme“, in: *Zeitschrift für experimentelle und angewandte Psychologie* (16), 157-183.

Toporov, V.N. (Hrsg.) (2005): *Jazyk – ličnost' – tekst. Sbornik v čest' T.M. Nikolaevoj*. Moskva.

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

Wimmer, G.; Altmann, G. (2005): „Unified Derivation of Some Linguistic Laws.“ In: Köhler/Altmann/Piotrowski (Hrsg.) (2005), 791-807.

Wimmer, G.; Altmann, G. (2006): „Towards a Unified Derivation of Some Linguistic Laws.“ In: Grzybek (Hrsg.) (2006), 329-337.

Kritik und Phrase

Festschrift für
Wolfgang Eismann
zum 65. Geburtstag

Herausgegeben von
Peter Deutschmann

unter Mitarbeit von
Peter Grzybek, Ludwig Karničar, Heinrich Pfandl

SONDERDRUCK

Praesens Verlag
Literaturwissenschaft | Sprachwissenschaft | Musikwissenschaft | Kulturwissenschaft
Wien

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

ISBN: 978-3-7069-0457-5

Gedruckt mit Förderung des Bundesministeriums
für Wissenschaft und Forschung in Wien

sowie des Landes Steiermark

und der Karl-Franzens-Universität Graz

© Praesens Verlag
http://www.praesens.at
Wien 2007
Alle Rechte vorbehalten. Rechtsinhaber, die nicht ermittelt werden
konnten, werden gebeten, sich an den Verlag zu wenden.

Inhalt

Wolfgang Eismann – Leben in Bewegung 11

Schriftenverzeichnis Wolfgang Eismann 23

“Kritik”

Dagmar Burkhart (Mannheim/Hamburg)
Viktor Pelevins *Schreckenshelm* als Relektüre des Mythos vom Minotaurus im Labyrinth 45

Peter Deutschmann (Graz)
Revizor revisited 71

Peter Grzybek (Graz)
What a Difference an »E« Makes: Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen 105

Emmerich Kelih (Graz)
Überlegungen zum sogenannten Kompensationsprinzip 129

Борис Успенский (Москва/Рим)
К истолкованию оды «Вольность» А. Н. Радищева 139

Christine Engel (Innsbruck)
Ein ewiger Stern im patriotischen Diskurs: *Zvezda* 1947-2002 143

Barbara Stelzl-Marx (Graz)
Čistoe nebo und *Sud’ba čeloveka*. Filmische Rehabilitierung sowjetischer Kriegsgefangener in der Tauwetterperiode 171

Dirk Kretzschmar (Szczecin)
Politik – multimedial. Der Internetauftritt des russischen Präsidenten Putin aus medientheoretischer und interkultureller Perspektive 189

Dagmar Gramshammer-Hohl (Graz)
Kritik der Differenz: Der Orientalismus als „blinder Fleck“ der Geschlechterforschung 213

Rainer Grübel (Oldenburg)
„Musik ist die Kurzschrift des Gefühls“ („Muzyka – ėto stenografija čuvstv“) 227

Andreas Guski (Basel)
Loyale Dissidenz. Über Sprache und Sprachkritik in Platonovs *Kotlovan* 249

Ingeborg Ohnheiser (Innsbruck)
Sprachgesetz und Sprachkritik. Zeitgenössische Schriftsteller/innen über jüngere Entwicklungen im Russischen 265

Наталия Брагина (Москва)
Интеллектуальные свойства памяти 285

Renate Hansen-Kokoruš (Mannheim)
Gore und *dolje* – kognitive, semantische und kulturelle Bedeutungsdimensionen in den südslawischen Sprachen 301

Walter Koschmal (Regensburg)
Die Rückkehr der Mitte. Zu einer Kritik des »Mitteleuropa«-Begriffs 319

Anke Levin-Steinmann (Leipzig)
Kritisches zu Schlüsselkonzepten, Stereotypen und Weltbildern 333

Klaus Trost (Regensburg)
Von Heidegger zu Koschmieder 355

“und”

Николай Алефиренко (Белгород)
Когнитивные основания лингвосемиозиса 379

Vladimir Biti (Zagreb)
Die eingebüßte Souveränität. Zur „Phraseologie“ der gegenwärtigen Theorie 397

Karl Eimermacher (Bochum/Berlin) / Valerij Gretchko (Bochum)
Aspekte einer interdisziplinären Untersuchungsstrategie im Bereich kultureller Evolution und Kreativität (Projektkonzept für ein künftiges Forschungszentrum) 405

Aage A. Hansen-Löve (München)
Die Natur der Natur bei Kazimir Malevič 423

Ingrid Schellbach-Kopra (München)
Aki Kaurismäki – Annäherung an ein Phänomen 441

Rupprecht S. Baur, Christoph Chlosta, Aneta Nickel (Essen)
Spätaussiedler – Identität und Sprache 459

Norbert Mappes-Niediek (Graz)
Nation an der Grenze 471

Heinrich Pfandl (Graz)
Kritik und nicht nur Phrasen: Friedrich Huldreich Erismann und seine Familie als Ausdruck einer gelebten Multikulturalität über drei Generationen 489

“Phrase”

Alexander Bierich (Heidelberg)
Polnisches und russisches Argot – ein Vergleich 511

Željka Fink (Zagreb)
Hrvatski frazemi kojima se opisuje neka ilegalna djelatnost (u usporedbi s ruskima) 523

Aleksandra Derganc (Ljubljana)
Revitalizacija ruskega frazema *и иже с ним (с ними)* 533

Maximilian Hendler (Graz)
Die FZ 5 im slowenischen Volksgesang und in der Welt 541

Ludwig Karničar (Graz)
Diatopische Synonyme für die Kartoffel in den Kärntner slowenischen Dialekten 553

Erika Kržišnik (Ljubljana)
Življenje nekega (kalkirano) prevzetega frazema 567

Josip Matešić (Mannheim)
O dijalektnoj frazeologiji u hrvatskome jeziku 579

Валерий Мокиенко (Санкт-Петербург)
Из истории фразеологических европеизмов: *Эй, завяжи на память узелок!* 585

Людмила Степанова (Olomouc)
Фразеологические заимствования в русском и чешском языках: старое и новое 597

Branko Tošović (Graz)
Phraseologische Unterschiede zwischen dem Bosnischen/Bosniakischen,
Kroatischen und Serbischen 611

Harry Walter (Greifswald)
Über phraseologische *Perlen* und eine *Mördergrube* 633

Олег Федосов (Будапешт)
Некоторые вопросы эквивалентности во фразеографии 643

Annelies Häcki Buhofer (Basel)
Phraseographie im Variantenwörterbuch des Deutschen 657

Vida Jesenšek (Maribor)
Idiomatik in allgemeinen slowenisch-deutschen Wörterbüchern.
Phraseologische und lexikographische Überlegungen 675

Wolfgang Stadler (Innsbruck)
Schweigephrasen – Schweigephasen 691

Harald Burger (Zürich)
Inszenierte Phraseologie – phraseologische Inszenierung: medienspezifische Aspekte des Phraseologie-Gebrauchs 707

Werner Helmich (Graz)
Die Entidiomatisierung von phraseologischem Material als Pointierungstechnik in französischen und deutschen Aphorismen 735

Anna T. Litovkina (Szekszárd)
“He Who Laughs, Lasts”: The Most Common Types of Alteration in Anglo-American Anti-Proverbs 761

Wolfgang Mieder (Burlington/Vermont)
„Ehrlich währt am kürzesten“. Phrasenkritik durch Antisprichwörter im 19. Jahrhundert 781

Manfred Schruba (Bochum)
Sentenzen und Sprichwörter in der Theorie und Praxis der russischen Komödie des 18. Jahrhunderts (Sumarokov, Nikolev, Plavil’ščikov) 801

Виктория Васильева (Пермь)
Традиционная паремия в рекламном слогане: тенденции и отношения 819