have a different proportion and therefore preference for the use of borrowed words, a finding that will contribute to applications in automatic text classification and genre detection, a promising potential that we are currently investigating in a separate study.

Acknowledgements

Research described in this article was supported in part by grants received from City University of Hong Kong (Project Nos 9610126, 7008002, 7002387 and 7002190).

References

ACL/DCI. (1993). *Linguistic Data Consortium.* Philadelphia.

Bar-Ilan, L., – Berman, R. A. (2007). Developing Register Differentiation: The Latinate-Germanic Divide in English". *Linguistics, 45(1), 1 – 35.*

The British National Corpus, Version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Burnard, L. (2007). *Reference Guide for the British National Corpus* (XML Edition). URL: http://www.natcorp.ox.ac.uk/XMLedition/URG/.

Culpeper, J., – Clapham, P. (1996). The Borrowing of Classical and Romance Words into English: A Study based on the Electronic Oxford English Dictionary. *International Journal of Corpus Linguistics, 1(2), 199 – 218.*

De Forest, M., – Johnson, E. (2001). The Density of Latinate Words in the Speeches of Jane Austen's Characters. *Literary and Linguistic Computing, 16(4), 389 – 401.*

Fang, A.C., – Cao, J., – Song, Y. (2009). A New Corpus Resource for Studies in the Syntactic Characteristics of Terminologies in Contemporary English. In: *Proceedings of the 8th Terminology and Artificial Intelligence Conference* (TIA'09), Toulouse, France.

Gramley, S., – Paetzold, K.-M. (1992). *A Survey of Modern English.* London and New York: Routledge.

Heylighen, F., – Dewaele, J.-M. (1999). *Formality of Language: Definition, Measurement and Behavioral Determinants.* Internal Report, Center "Leo Apostel", Free University of Brussels. Available at: http://pespmc1.vub.ac.be/papers/Formality.pdf

Hoffmann, S. (2009). Lexical Change. In: J. Culpeper, F. Katamba, P. Kerwill, R. Wodak & T. McEnery (EDs*), English Language: Description, Variation and Context*, Palgrave MacMillan, 286 – 300.

Laar, M. (1998). The Latin Component in English Medical Texts and Some of the Possibilities it Offers for Interdisciplinary Integrated Teaching. In: *Proceedings of Linguistics in Estonia and Finland: Crossing the Fulf Symposium, Tallinn, Estonia, 171 – 174.*

Lee, D. (2001). Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning and Technology, 5(3), 37 – 72.*

Márquez, M.F. (2007). Renewal of Core English Vocabulary: A Study Based on the BNC. *English Studies, 88(6), 699 – 723.*

Peters, P. (1998). Surveying Contemporary English Usage. *English Today, 56, 14(4), 3 – 6.*

Roberts, H.A. (1965). A Statistical Linguistic Analysis of American English. The Hague: Mouton.

Stockwell, R. – Minkova, D. (2001). *English Words: History and Structure.* Cambridge: Cambridge University Press.

GLOTTOTHEORY 2010, NUMBER 3/1, PP 69 – 82.

Kumulierte Ranghäufigkeiten von slawischen Graphemen: Modell und Parameter-Interpretation

Emmerich Kelih[1]

Abstract. *In the article one particular continuous function is used to fit cumulative grapheme rank frequencies from a parallel corpus. Furthermore some interrelations between the repeat rate, first rank frequency, mean and some theoretical parameters are shown.*

0 Einleitung

Die Modellierung von Graphemranghäufigkeiten stand in letzter Zeit immer wieder im Mittelpunkt des Interesses der quantitativen Linguistik. Insbesondere geht es um die Frage eines statistischen Modells, welches in der Lage ist Graphem- bzw. Phonemhäufigkeiten adäquat zu beschreiben. Einerseits erfolgt dies durch diskrete Verteilungsmodelle (vgl. Best 2005; Grzybek, Kelih, Altmann (2004), Grzybek, Kelih 2005a u.v.m.) und andererseits wurden in Kelih (2009a) auf der Basis von slawischen Paralleltexten mehrere, in der Vergangenheit an unterschiedlicher Stelle vorgeschlagene, stetige Funktionen einer eingehenden empirischen Überprüfung unterzogen.

Auch wenn bislang aus dieser Reihe von Untersuchungen noch kein Modell aufgrund der Anpassungsgüte, oder aber auch aufgrund eines interpretierbaren Verhaltenes der in diese Modelle einfließenden Parameter, favorisiert werden kann, soll im folgenden Beitrag auf eine weiterführende Frage eingegangen werden. Diese besteht darin, auf welche Art und Weise kumulierte Ranghäufigkeiten von Graphemhäufigkeiten statistisch erfasst werden können. Diese Problemstellung wurde bislang nicht systematisch untersucht. Folgende Probleme stehen zur Diskussion: (1) Die Diskussion einer geeigneten stetigen Funktion für kumulierte Ranghäufigkeiten von Graphemen und (2) das Aufzeigen von Wechselbeziehungen zwischen aus den Rangverteilungen berechneten empirischen Kenngrößen (Inventarumfang, Mittelwert, Repeat-Rate und die erste Häufigkeitsklasse (p_1)) und den in das Modell eingehenden theoretischen Parametern. Ziel ist es auf diese Art und Weise die entsprechenden theoretischen Parameter linguistisch interpretieren zu können.

1 Kumulierte Ranghäufigkeiten

Die kumulierten Ranghäufigkeiten von Graphem- bzw. Phonemhäufigkeiten waren bislang kaum Gegenstand von systematischen Ausführungen. Das prinzipielle Vorgehen soll anhand eines in 1.2 näher beschriebenen Datensatzes zum Slowenischen näher beschrieben werden. Als Ausgangsbasis dienen Graphemhäufigkeiten in Texten, die in eine Ranghäufigkeit transformiert werden. In einem nächsten Schritt werden die absoluten Häufigkeiten in relative Häufigkeiten überführt und sodann wird ausgehend von der ersten Klasse eine kumulative Ranghäufigkeit gebildet.

[1] Institut für Slawistik, Universität Graz, Merangasse 70/1, 8010 Graz, Austria. emmerich.kelih@uni-graz.at

Tabelle 1: Rohdaten für Slowenisch

Rang	abs. F.	rel. F.	kum. F.
1	30849	0,1068	0,1068
2	29708	0,1028	0,2096
3	26129	0,0905	0,3001
4	25886	0,0896	0,3897
5	17175	0,0595	0,4492
6	15921	0,0551	0,5043
7	15045	0,0521	0,5563
8	14144	0,0490	0,6053
9	14139	0,0489	0,6543
10	12402	0,0429	0,6972
11	11569	0,0400	0,7372
12	11412	0,0395	0,7767
13	10029	0,0347	0,8115
14	9167	0,0317	0,8432
15	8753	0,0303	0,8735
16	6441	0,0223	0,8958
17	5515	0,0191	0,9149
18	5336	0,0185	0,9334
19	4755	0,0165	0,9498
20	4429	0,0153	0,9652
21	3054	0,0106	0,9757
22	2923	0,0101	0,9858
23	1967	0,0068	0,9927
24	1893	0,0066	0,9992
25	230	0,0008	1
Summe	288871	1	

Wie die Gegenüberstellung der relativen Ranghäufigkeiten mit den kumulierten Ranghäufigkeiten in Abb. 1a. und Abb.1b zeigt, ist die Kumulierung als eine Möglichkeit zu sehen um – wie in der Graphik zu sehen, – auftretende Sprünge in einer „normalen" Rangverteilung in gewisser Weise zu „glätten". Die Gründe für das Auftreten dieser Sprünge sind bislang nicht bekannt. Zu vermuten ist, dass es sich hierbei um unterschiedliche Strata handelt (vgl. Popescu, Altmann, Köhler 2010).

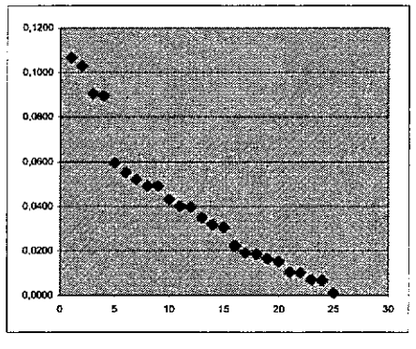

Abb.1a. Relative Ranghäufigkeiten (Slowenisch)

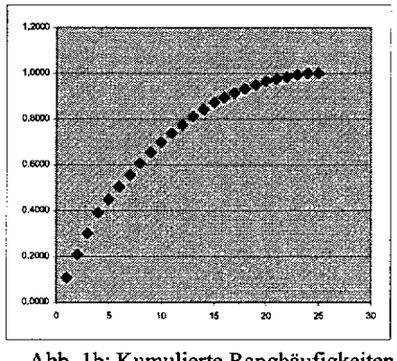

Abb. 1b: Kumulierte Ranghäufigkeiten (Slowenisch)

Es zeigt sich deutlich, dass die kumulierten Ranghäufigkeiten insgesamt einen mehr „glatten" Anstieg aufweisen als die nicht kumulierten Häufigkeiten. Es ist anzunehmen, dass bei der theoretischen Modellbildung der kumulierten Ranghäufigkeiten von anderen, möglicherweise einfacheren Voraussetzungen auszugehen ist, als bei der Modellierung der üblichen Rangverteilungen von Graphemhäufigkeiten. Darüber hinaus wird zu klären sein, welche linguistische Aussagekraft kumulierte Ranghäufigkeiten haben.

1.1 Stetige Modelle: Theorie

Die linguistische Bedeutung einer kumulierten Ranghäufigkeit liegt darin, dass diese Art der Dateninterpretation eine Möglichkeit darstellt, um die Dynamik des Anwachsens von Graphemen in einem Text untersuchen zu können. Es ist bislang weitgehend unbekannt, von welchen Faktoren die Sättigung eines Textes durch Grapheme abhängt. Als direkter Einflussfaktor ist vor allem der Inventarumfang von Graphemen/Phonemen zu nennen. Es ist plausibel, dass z.B. Sprachen mit kleinen Inventaren einen schnelleren Anstieg der kumulierten Häufigkeiten aufweisen, als Sprachen mit größeren Inventarumfängen.

Eine Untersuchung derartiger Fragestellungen impliziert das Finden eines theoretischen Modells, welches in der Lage ist, die kumulierten Ranghäufigkeiten adäquat zu erfassen. Besondere Aufmerksamkeit ist dabei darauf zu legen, dass die „Zunahmegeschwindigkeit" der Grapheme zu einem zentralen Bestandteil des Modells wird. Zu beachten ist auch, dass vor allem im vorderen Rangbereich eine schnelle Zunahme, d.h. hohe Sättigung, empirisch zu beobachten ist.

In Anlehnung an Popescu et al. (2009: 101f), die die Kumulationen von lexikalischen Einheiten untersucht haben, soll ausgehend von der Differentialgleichung

$$(1) \qquad \frac{dy}{dx} = g(a-y)\ ,$$

in der g die Proportionalitätskonstante und a die Asymptote darstellt, die entsprechende Sättigungsgeschwindigkeit von Graphemhäufigkeiten erfasst werden. Löst man (1) auf und setzt $y = 0$, so ergibt sich

$$(2) \qquad y = a(1-\exp(-gx))\,.$$

Man hat es somit mit einer Exponentialfunktion mit zwei Parametern a und g zu tun, die in der Lage sein sollte das Anwachsen von Graphemen zu erfassen.

Die Übertragung dieses Modells, welches sich bislang für die Modellierung lexikalischer Einheiten bewährt hat, auf kumulierte Ranghäufigkeiten von Graphemen ist zwar nicht unproblematisch, kann aber theoretisch damit begründet werden, dass für sprachliche Prozesse sehr allgemeine, aber für unterschiedliche Ebenen ähnliche Sättigungsprozesse anzunehmen sind. Ob das vorgeschlagene Modell in der Tat auch für kumulierte Ranghäufigkeiten von Graphemen geeignet ist, wird im Folgenden empirisch zu überprüfen sein.

1.2 Empirische Ergebnisse: Ranghäufigkeiten

Das in Kap. 1.1. vorgestellte Modell wird nunmehr auf der Basis von elf slawischen Sprachen empirisch getestet. Als Datenbasis dient ein vom Autor erstelltes Korpus von Übersetzungen des russischen Romans von N. Ostrovskij „Kak zakaljalas' stal' // Wie der Stahl gehärtet wurde" (1932 – 1934). Das Paralleltext-Korpus ist in Kelih (2009b und 2009c) ausführlich

beschrieben und bereits hinsichtlich der Anzahl von Types und Tokens auch in Ansätzen statistisch untersucht. In Kelih (2009a) sind die Rohdaten (Graphemhäufigkeiten) zu allen elf[2] untersuchten slawischen Sprachen und dem russischen Originaltext abgedruckt. Im Anhang 1 dieser Arbeit sind daher nur mehr die entsprechenden kumulierten relativen Ranghäufigkeiten für alle untersuchten Sprachen abgedruckt.

Hinsichtlich des in Tabelle 2 angeführten Inventarumfanges k sei spezifiziert, dass es sich hierbei um jenen Inventarumfang handelt, der sich aus der Anzahl von empirisch realisierten Einheiten ergibt. D.h. es wurde jener Inventarumfang pro Sprache herangezogen, der in dem Text realisiert ist, und nicht jener Inventarumfang, der gegebenenfalls auf paradigmatischer Ebene, wie beispielsweise in Referenzgrammatiken, postuliert wird. Als Anpassungsgüte wird der Determinationskoeffizient R^2 verwendet. Ein $R^2 > 0.85$ wird als passable Übereinstimmung zwischen empirischen und theoretischen Werten gewertet.

Die Ergebnisse zur Modellierung der kumulierten Ranghäufigkeiten zeigen, dass sich in fast allen Sprachen ein $R^2 > 0.99$ (nur für Mazedonisch $R^2 = 0.9881$) ergibt, was insgesamt als vortreffliches Resultat verstanden werden kann.

Tabelle 2[3]: Anpassungsergebnisse und Parameter a und g

Sprache	Inventarumfang k	a	g	R^2
Slo	25	0,1254	1,0241	0,9901
Serb	30	1,0533	0,1084	0,9978
Cro	30	1,0563	0,1070	0,9976
Bulg	30	1,0406	0,1130	0,9993
Mz	31	0,1035	1,0938	0,9881
Rus	32	1,0666	0,0949	0,9994
Ukr	34	1,0839	0,0840	0,9987
Cz	41	1,0639	0,0805	0,9985
Sk	42	0,0820	1,0750	0,9940
Pl	39	1,0943	0,0842	0,9988
O-Srb	34	1,0656	0,0841	0,9980

Damit zeigt sich, dass ein passendes Modell für die Modellierung der kumulativen Ranghäufigkeiten von Graphemen gefunden worden ist. Zu hoffen bleibt, dass sich dieses Modell auch für weitere, vor allem auch nichtslawische Sprachen als adäquat erweisen wird.

1.3 Empirische Kenngrößen von Rangverteilungen

Die Modellierung ist ein erster Zwischenschritt der systematischen Untersuchung von kumulierten Ranghäufigkeiten. Von besonderem Interesse ist nun die Frage, inwiefern und ob

[2] Das Weißrussische wird hier von der Untersuchung ausgeschlossen. Dies ist dadurch zu begründen, dass, wie in Kelih (2009b) gezeigt, die Graphemhäufigkeiten dieser Sprache hinsichtlich von theoretischen Modellen nicht dem Gesamtbild in allen anderen slawischen Sprachen entspricht. Linguistisch wird dieses Abweichen dadurch begründet, dass das Weißrussische über eine phonetisch geprägte Orthographie verfügt.

[3] Für die Sprachen werden in den Tabellen und den Abbildungen folgende Abkürzungen eingeführt: Bulg = Bulgarisch, Kro = Kroatisch, Mz = Mazedonisch, O-Srb = Obersorbisch, Pl = Polnisch, Rus = Russisch, Serb = Serbisch, Sk = Slowakisch, Slo = Slowenisch, Tsch = Tschechisch, Ukr = Ukrainisch.

die Proportionalitätskonstante g interpretierbar ist und in welchem Zusammenhang sie zu ausgewählten empirischen Kenngrößen, wie dem Mittelwert, der Wiederholungsrate, der ersten Häufigkeitsklasse p_1 und dem Inventarumfang k steht.

Folgende statistische Kenngrößen stehen unter anderem zur Verfügung, um eine empirische Rangverteilung adäquat beschreiben zu können:

(1) Erste wichtige Kenngröße ist der Mittelwert $\bar{x}$, der sich als $\bar{x} = \frac{1}{N}\sum_{x=1}^{k} xf(x)$ berechnen lässt, wo k der Inventarumfang, N die Anzahl aller Häufigkeiten und $f(x)$ die empirischen Häufigkeiten bedeutet. Der Mittelwert ist ein zentrales „Lokationsmaß" einer Rangverteilung.

(2) Die Spannweite der Rangverteilung, d.h. der zugrunde gelegte Inventarumfang (k) ist eine wichtige Steuerungsgröße und hat einen direkten Einfluss auf die Form einer Ranghäufigkeit. Darüber hinaus ist die Silben-, Morphem- und Wortstruktur durch den Inventarumfang einer Sprache beeinflusst (vgl. u.a. Strauss, Fan, Altmann 2008: 8, 105, und 113). Der Inventarumfang einer Sprache kann auf unterschiedliche Weisen bestimmt werden und hängt letztlich von der Zugehörigkeit zu einer bestimmten linguistischen Schule ab. Zu beachten wäre allerdings in jedem Fall, dass sprachliche Systeme in vielen Fällen über so etwas wie „periphere" Grapheme verfügen. Das sind Grapheme, die kaum in ein sprachliches System integriert sind. Dies kann z.B. der Fall sein, wenn bestimmte Grapheme/Phoneme nur in Fremdwörtern gebraucht werden, oder aber die Phoneme, die durch diese Grapheme ausgedrückt werden, bereits im Laufe ihrer Sprachgeschichte ihre distinktive Kraft verloren haben. Dieser Sachverhalt bewirkt aber, dass der untere Teil einer Rangverteilung marginal besetzt ist. Dies zeigt sich auch daran, dass unter Umständen die empirischen Rangverteilungen von Graphemhäufigkeiten zwar im vorderen Rangbereich stark „besetzt" sind, aber danach ein schnelles „Abflachen" zu beobachten ist und darüber hinaus einzelne Grapheme sich überhaupt nicht in einen sich abzeichnenden Gesamttrend integrieren lassen.

(3) Als linguistischer Input wird die erste Ranghäufigkeit (p_1) eingeführt, d.h. jene Häufigkeit, die in einem sprachlichen System am häufigsten genützt wird und darüber hinaus ist p_1 eine Art „Startwert" einer Ranghäufigkeitsverteilung.

(4) Weiteres wichtiges Maß ist die (in der quantitativen Phonologie) häufig diskutierte (vgl. Altmann, Lehfeldt 1980: 151f.) Wiederholungsrate[4] (R). Diese lässt sich als $R = \sum_{r=1}^{k} p_r^2$ berechnen, d.h. als die Summe der quadrierten relativen Häufigkeiten. Zur Wiederholungsrate und Entropie im Russischen vgl. Grzybek, Kelih, Altmann (2005: 122f.). Die Wiederholungsrate ist ein Maß der Gleichverteilung und gleichzeitig eine wichtige systemlinguistische Eigenschaft.

[4] Die Wiederholungsrate ist ein Maß der Gleichverteilung und es ist gut bekannt (vgl. Altmann, Lehfeldt 1980), dass die Wiederholungsrate und der Inventarumfang von Graphem- bzw. Phonemhäufigkeiten in einer gegenseitigen Wechselbeziehung stehen. So wird in Altmann, Lehfeldt (1980: 158f.) auf der Basis von 63 Sprachen festgestellt, dass mit zunehmenden Inventarumfang der Sprachen die Wiederholungsrate sinkt. Das heißt, vereinfacht gesagt, dass Sprachen mit einem geringen Inventar die Tendenz zeigen, bestimmte Phoneme sehr stark auszulasten, während ein hoher Inventarumfang eher eine Gleichverteilung der Phoneme zu begünstigen scheint. Dieser Zusammenhang wird in Altmann, Lehfeldt (1980: 171f) mit der Potenzfunktion $y = ax^{-b}$ erfasst. In einer Re-Analyse durch Grzybek, Kelih (2005) wurde der Determinationskoeffizient berechnet, der $R^2 = 0.7692$ beträgt; für weitere Details vgl. Grzybek, Kelih (2005: 66).

Damit liegen insgesamt vier Merkmale von Graphemhäufigkeiten vor, deren Zusammenhänge zur Proportionalitätskonstante g nun zu klären sein werden. Die Werte für die einzelnen Sprachen sind in Tabelle 3 zusammengefasst.

Tab. 3: Empirische Kennwerte von Graphemranghäufigkeiten

Sprachen	k	p_1	$\bar{x}$	R
Slo	25	0,1068	7,87	0,0624
Serb	30	0,1225	8,21	0,0621
Cro	30	0,1204	8,25	0,0616
Bulg	30	0,1334	8,15	0,0627
Mz	31	0,1419	7,46	0,0696
Rus	32	0,1064	9,06	0,0543
Ukr	34	0,0965	9,81	0,0495
Czech	41	0,0806	10,62	0,0455
SK	42	0,1028	9,84	0,0507
Poln.	39	0,0915	10,42	0,0489
O-Sorb.	34	0,0988	10,15	0,0494

1.3.1 Mittelwert und Wiederholungsrate

Bevor auf die Bedeutung der Proportionalitätskonstante g des theoretischen Modells eingegangen werden kann, ist festzuhalten, dass es auch zwischen den erwähnten empirischen Kenngrößen unterschiedlich starke statistische Zusammenhänge gibt. Auf diese Wechselbeziehungen sollte bereits im Vorfeld jeglicher weiterführenden Analyse und Interpretation hingewiesen werden.

Zu beginnen ist mit dem Zusammenhang zwischen dem Mittelwert und der Wiederholungsrate: Je höher der Mittelwert einer Rangverteilung, desto niedriger die Wiederholungsrate. Diese Hypothese lässt sich dadurch begründen, dass ein hoher Mittelwert dazu führt, dass Grapheme gleichmäßiger ausgenützt werden, und somit die Wiederholungsrate ebenfalls niedrig gehalten wird. D.h. zwischen dem Mittelwert einer Rangverteilung und der Wiederholungsrate müsste eine derartige, angeführte gesetzesartige Beziehung bestehen.

Wie eine empirische Überprüfung zeigt, kann in der Tat folgendes einfaches Potenzmodell als adäquat für die Beschreibung dieses Zusammenhanges angesehen werden: $R = 0.62\bar{x}^{-1.097}$. Es ergibt sich ein $R^2 = 0.9801$. Dieser Wert ist als ein hervorragendes Ergebnis zu interpretieren. Vgl. dazu auch Abb.2.

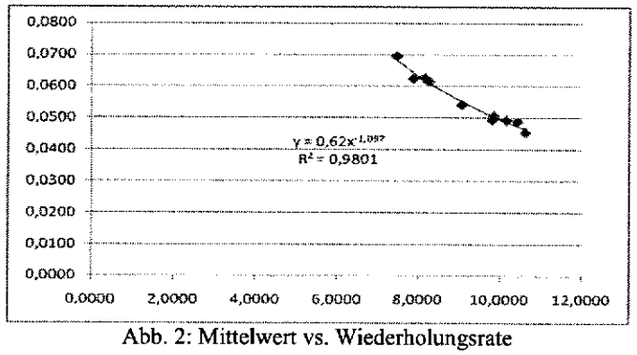

Abb. 2: Mittelwert vs. Wiederholungsrate

Dieser Befund bedeutet aber auch gleichzeitig, dass hinsichtlich eines Zusammenhanges zwischen der Proportionalitätskonstante g und dem Mittelwert bzw. der Wiederholungsrate R nur eine der Kenngrößen im Detail interpretiert werden muss, denn der Mittelwert und die Wiederholungsrate verfügen, wie soeben gezeigt, im vorliegenden Fall über den gleichen Informationswert.

1.3.2 Wiederholungsrate vs. p_1

Ähnliches – wie soeben festgestellt – gilt auch für einen Zusammenhang zwischen der Wiederholungsrate und der ersten Häufigkeitsklasse (p_1). Auch in diesem Fall kann ein hoher statistischer Zusammenhang zwischen diesen beiden Variablen festgestellt werden, denn es kann postuliert werden, dass je höher p_1, desto höher auch die Wiederholungsrate. Dies lässt sich dadurch erklären, dass ein hoher Redundanzgrad (d.h. ein hohes p_1) mit einer hohen Wiederholungsrate einhergeht.

Folgendes Modell in der Form von $R = 0.4032{p_1}^{0.0119}$ beschreibt diese Zusammenhänge recht gut, wird doch ein $R^2 = 0.8622$ erreicht. Auch wenn somit die Stärke (vgl. dazu auch Abb. 3) des Zusammenhanges etwas schwächer ausgeprägt ist als im Fall des Zusammenhanges zwischen dem Mittelwert und der Repeat-Rate, kann die Anpassung als zufriedenstellend akzeptiert werden.

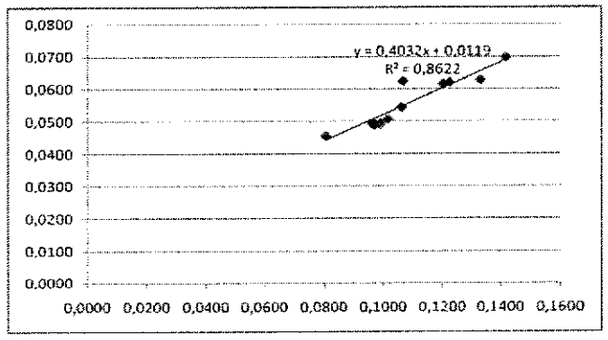

Abb. 3: p_1 vs. Wiederholungsrate

Nachdem nun einige „interne" Zusammenhänge (d.h. zwischen den empirischen Kenngrößen) festgestellt werden konnten, sollen im Folgenden deren Zusammenhänge zur Proportionalitätskonstante g des oben angeführten theoretischen Modells näher untersucht werden.

1.3.3 Interpretation der Parameter

Aus linguistischer Sicht kann der Verlauf des Anwachsens von kumulierten Ranggraphemhäufigkeiten von unterschiedlichen Faktoren abhängen: Einerseits ist zu vermuten, dass p_1 im Sinne eines „Startwertes" eine hervorragende Rolle spielt, indem er gleichsam den Verlauf, insbesondere aber den vorderen Rangbereich, in gewisser Weise „vorwegnimmt" bzw. determiniert. Eine ähnliche Bedeutung könnte auch der Wiederholungsrate zukommen, da eine ungleiche Verteilung von Phonem- und Graphemhäufigkeiten entweder zu einem raschen oder aber verlangsamten Anstieg der Verteilungskurve führen kann.

Und, wie bereits erwähnt, sollte allerdings der Inventarumfang als direkt interpretierbare linguistische Größe einen wichtigen Einfluss auf die Anstiegsgeschwindigkeit der kumu-

lierten Ranghäufigkeiten von Graphemen haben: Ein kleines Phonem- bzw. Grapheminventar müsste ein schnelleres Anwachsen nach sich ziehen als ein Inventar mit vielen Einheiten. Linguistisch spielt der Inventarumfang eine zentrale Rolle, nicht nur in Bezug auf Graphemhäufigkeiten, sondern vor allem auch hinsichtlich seiner „determinierenden Funktion" auf sprachlich gesehen höheren Ebenen, wie in etwa die Silben-, Morphem- und Wortstruktur. Der bekannteste Zusammenhang, der hier allerdings nicht im Detail vorgestellt werden muss (vgl. Kelih 2008), bezieht sich auf den Zusammenhang zwischen dem Inventarumfang von Phonemen/Graphemen und der Wortlänge.

Betrachtet man Abb. 4, so zeigt sich tatsächlich, dass hinsichtlich eines Zusammenhanges von Inventarumfang und Proportionalitätskonstante *g* auf der Basis des vorliegenden Materials in der Tat von einer gesetzesartigen Beziehung gesprochen werden kann.

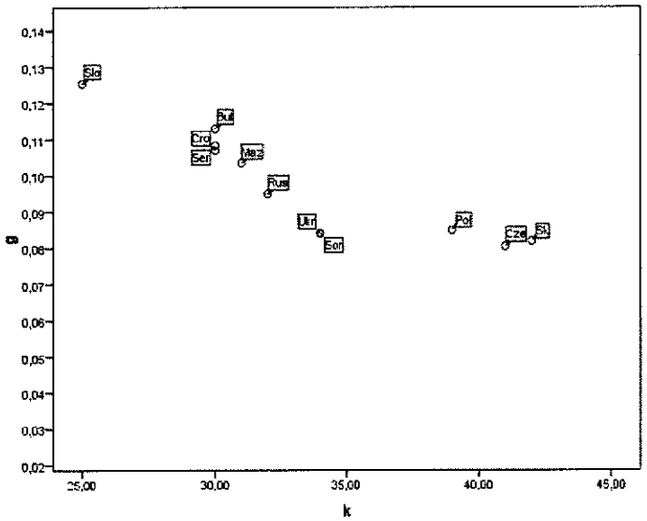

Abb. 4: Inventarumfang vs. Parameter *g*

Es lässt sich ein Trend und linguistisch begründbares Phänomen erkennen, wonach mit zunehmendem Inventarumfang der Parameter *g* sinkt. Versucht man den Zusammenhang mit Hilfe eines statistischen Modell in der Form von $g = 2.2764*k^{-0.905}$ so ergibt sich ein $R^2 = 0.8435$. Vgl. dazu auch Abb. 5 mit einer entsprechenden Darstellung der Anpassungskurve.

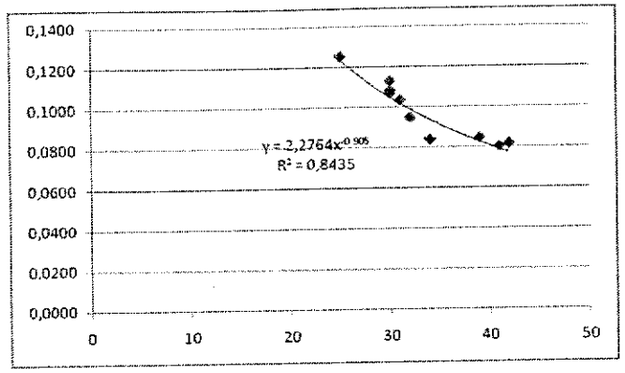

Abb. 7: Inventarumfang vs. Parameter *g*: Empirische Werte und theoretische Kurve

Somit lässt sich anhand des vorliegenden Datenmaterials bestätigen, dass der Inventarumfang als eine beträchtliche Einflussgröße von kumulierten Graphemranghäufigkeiten zu bezeichnen ist. Der Inventarumfang ist offenbar in der Lage die Geschwindigkeit des Anwachsens der kumulierten Ranghäufigkeiten in entsprechender Weise zu steuern. In zukünftigen Arbeiten wird zu prüfen sein, ob auch Ähnliches auf Ebene von nicht kumulierten Ranghäufigkeiten zu beobachten ist.

Abschließend sind die bislang gefundenen Zusammenhänge zwischen den empirischen Kenngrößen (p_1, R, k) und dem Parameter *g* aus dem theoretischen Verteilungsmodell für kumulierte Ranghäufigkeiten in Form eines einfachen synergetischen Regelkreises [5] darzustellen.

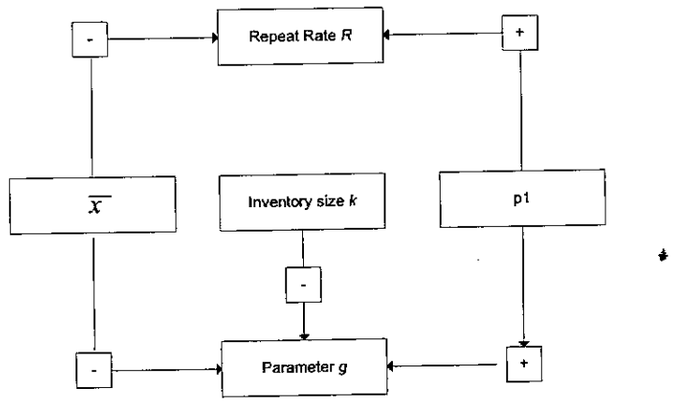

Abb. 8: Regelkreis von empirischen Kenngrößen und theoretischem Parameter *g*

Abschließend ist noch zu dem von uns gewählten Modell zu sagen: Betrachtet man die Funktion (2) als die (kumulative) Verteilungsfunktion der Ranghäufigkeiten $F(r)$, dann ergibt die Ableitung $F'(r) = f(r)$ genau die erste Komponente der stratifizierten Variante des

[5] Die Überprüfung des Zusammenhanges von Mittelwert und Parameter *g* ergab sich bei Verwendung von $g = 1{,}1067\bar{x}^{-1{,}112}$ ein gutes $R^2 = 0.8247$. Für $g = 0.4275p_1^{0.6719}$ ergibt sich allerdings ein nicht allzu überzeugendes von $R^2 = 0{,}5127$, sodass wohl einstweilen in diesem Zusammenhang von einem tendenziellen Zusammenhang gesprochen werden sollte.

Zipfschen Gesetzes, so wie sie von Popescu, Altmann und Köhler (2010) vorgeschlagen wurde. Die Sprünge zwischen den einzelnen Rängen, wie in Abb. 1.a sichtbar, weisen genau auf diese Art von Stratifikation hin, die sich mit weiteren Komponenten überbrücken ließe.

2 Zusammenfassung

Fassen wir zusammen: Erstes wichtiges Ergebnis ist, dass ein passendes – ursprünglich in der Lexik diskutiertes und angewandtes – Modell für die Modellierung kumulierter Ranghäufigkeiten von Graphemen in elf slawischen Sprachen gefunden werden konnte. Damit scheint sich zu bestätigen, dass die Sättigung bzw. die Dynamik des Zuwachses von sprachlichen Einheiten durch ein und denselben statistischen Mechanismus gesteuert wird.

Als weiteres Ergebnis ergibt sich, dass sowohl Mittelwert und Wiederholungsrate von Graphemranghäufigkeiten als auch die erste Häufigkeitsklasse p_1 und die Wiederholungsrate R direkt voneinander und miteinander abhängen. Diese als intern zu bezeichnenden Abhängigkeiten sind insofern von Interesse, da sie die Möglichkeit einer Reduktion von zu interpretierenden Einflussfaktoren ermöglichen.

Hinsichtlich der Proportionalitätskonstante g des theoretischen Modells lässt sich festhalten, dass diese Größe einen starken Zusammenhang mit dem Inventarumfang aufweist. Diese lässt sich durch ein einfaches statistisches Modell erfassen. Damit zeichnet sich insgesamt ein komplexer Regelkreis ab, der aus „normalen" Ranghäufigkeiten und den kumulierten Ranghäufigkeiten abgeleitet werden kann. In Zukunft wird ein derartiger Regelkreis unter Hinzuziehung weiterer Kenngrößen (wie dem h-Punkt, der Bogenlänge usw.) zu erarbeiten sein, der sowohl das Verhalten von kumulierten als auch nicht kumulierten Ranghäufigkeiten von Graphemen/ Phonemen umfasst.

Literatur

Altmann, G.; Lehfeldt, W. (1980): *Einführung in die Quantitative Phonologie.* Bochum: Brockmeyer. [= Quantitative Linguistics, 7]

Best, K.-H. (2005). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovij visnik Černives'koho universitetu, vypusk (Hermans'ka filolohija) 231, 119 – 127.*

Grzybek, P. – Kelih, E. (2005a). Towards a general model of grapheme frequencies in Slavic languages. In: Garabík, R. (ed.), *Computer Treatment of Slavic and East European Languages: 73 – 87.* Bratislava: Veda.

Grzybek, P. – Kelih, E. (2005b). Häufigkeiten von Buchstaben/Graphemen/Phonemen: Konvergenzen des Rangierungsverhaltens. *Glottometrics 9, 62 – 73.*

Grzybek, P. – Kelih, E. – Altmann, G. (2004). Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie 32, 25 – 54.*

Grzybek, P. – Kelih, E. – Altmann, G. (2005). Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das 'ë'. *Anzeiger für Slavische Philologie 33, 117 – 140.*

Kelih, E. (2009a): Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle, in: *Glottometrics 18, 53 – 69.*

Kelih, E. (2009b): Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In: Kelih, E.; Levickij, V.V.; Altmann, G. (Hg.): Methods of Text Analysis. Metody analizu tekstu. Černivci: ČNU, S. 106–124.

Kelih, E. (2009c): Preliminary analysis of a Slavic parallel corpus. In: Levická, J.; Garabík, R. (eds.) (2009): *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25 – 27 November 2009. Proceedings.* Bratislava: Tribun, 175 – 183.

Popescu, I.-I. in co-operation with **Altmann, G.; Grzybek, P.; Jayaram, B.D.; Köhler, R.; Krupa,V.; Mačutek, J.; Pustet, R.; Uhlířová, L.; Vidya, M.N.** (2009): *Word frequency studies.* Mouton de Gruyter: Berlin – New York. [= Quantitative Linguistics, 64]

Popescu, I.-I. – Altmann, G. – Köhler, R. (2010). Zipf's law – another view. *Quality and Quantitty 44, 713 – 731* (DOI 10.1007/s11135-009-9234-y)