Glottometrics 9, 2005, 62-73

Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens

Peter Grzybek, Graz[1]
Emmerich Kelih, Graz[2]

Abstract. The present study raises the question in how far low-level linguistic units, such as letters, graphemes, sounds and phonemes, follow one and the same pattern as to their frequency distribution. Based on Altmann/Lehfeldt's (1980) study on 63 samples from 38 different languages, a separate re-analysis of the letter/grapheme vs. sound/phoneme samples is made, concentrating on the empirical entropy and repeat rate, on the one hand, and their theoretical calculations derived from the geometric and Zipf-Mandelbrot distributions. As a result, there are no significant differences as to these two global measures. This finding is interpreted in terms of a strong argument in favor of an analogical behavior of these linguistic units.

Keywords: letter frequencies, grapheme frequencies, phoneme frequencies, entropy, repeat rate, geometric distribution, Zipf-Mandelbrot distribution

1. Einführung

In der vorliegenden Untersuchung geht es um die Frage, ob sich die „niedrigsten" Spracheinheiten – nämlich Phoneme, Buchstaben, Grapheme – hinsichtlich ihrer Häufigkeitsverteilung analog verhalten, d.h. ob man für ihre Erfassung die gleichen Modelle benutzen kann. Hintergrund dieser Fragestellung ist der Umstand, dass Untersuchungen zur Vorkommenshäufigkeit von Buchstaben bzw. Graphemen in der jüngsten Zeit eine gewisse Renaissance zu erleben scheinen (s. z.B.: Best 2005a,b; Grzybek, Kelih, Altmann 2004, 2005a,b; Hussien 2004; Pääkkönen 1993; Rosenbaum, Fleischmann 2002, 2003); auch im Internet finden sich mittlerweile zu den verschiedensten Sprachen[3] Angaben zu Buchstabenhäufigkeiten, wobei allerdings nicht selten die einfachsten Voraussetzungen wissenschaftlichen Arbeitens verstoßen wird: So wird oft überhaupt nicht gesagt, auf welchem Material die jeweiligen Angaben beruhen, es werden keine absoluten, sondern nur relative Häufigkeiten angegeben (die für eine Reihe von weiterführenden Fragen unzureichend sind), usw. usf.

Beim Thema von Graphem- oder Phonemhäufigkeiten handelt es sich zweifellos um eine der traditionellsten Fragestellungen in der Geschichte der Quantitativen Linguistik, die in der Vergangenheit immer wieder mit den unterschiedlichsten pragmatischen Fragestellungen verbunden oder gar auf diese hin ausgerichtet war (vgl. Grzybek, Kelih 2003). So ging es in den wenigsten Untersuchungen um die einfache Erhebung von Buchstabenhäufigkeiten an und für

[1] Send correspondence to: Peter Grybek: peter.grzybek@uni-graz.at.
[2] Der Beitrag von Emmerich Kelih verdankt sich u.a. dem Doktorandenprogramm der Österreichischen Akademie der Wissenschaften (DOC).
[3] Auf den Versuch einer vollständigen Anführung von konkreten Angaben kann hier verzichtet werden: In der Regel handelt es sich hier um Internetseiten aus dem Bereich der Kryptographie bzw. Kryptologie; vgl. z.B.: http://www.cryptogram.org/cdb/words/frequency.txt#calcs, http://www.central.edu/homepages/LintonT/classes/spring01/cryptography/letterfreq.html

sich; vielmehr waren praktisch 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, die Relevanz ihrer Untersuchungen von Graphemhäufigkeiten für Fragen der Stenographie, der Tastaturbelegung von Schreibmaschinen bis hin zu Fragen der Textstilistik und Texttypologie, oder des Vergleichs mit der Phonologie.

Während in der Vergangenheit allerdings das Interesse bei der Untersuchung von Graphemhäufigkeiten in der Regel eher auf die relative Häufigkeit des Vorkommens einzelner Grapheme ausgerichtet war, wurde in jüngerer Zeit vermehrt die Frage nach einem allgemeinen Häufigkeitsmodell gestellt (Sigurd 1968; Orlov et al. 1982; Tuldava 1988; Naranan, Balasubrahmanyan 1992a,b; Martindale et al. 1996; Altmann 1993; Grzybek, Kelih, Altmann 2004, 2005a,b). Bei dieser Art von Fragestellung geht es weniger um die Frequenz der individuellen Buchstaben, sondern darum, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, zum dritthäufigsten, usw. hat. In den Vordergrund rückt damit eine Rang-Häufigkeitsverteilung, und das Ziel der theoretischen Modellierung ist die mathematische Formalisierung des Abstands zwischen den jeweiligen Häufigkeiten. Das Vorgehen hat man sich dabei wie folgt vorzustellen: Überführt man erhobene Ausgangsdaten in eine Rang-Reihenfolge, so geschieht dies üblicherweise in absteigender Reihenfolge. Wenn man sodann die jeweiligen Datenpunkte miteinander verbindet, ergibt sich charakteristischerweise kein linearer Abfall, sondern eine spezifische, monoton fallende (üblicherweise hyperbolische) Kurve. Und genau darum ist es in den genannten Untersuchungen gegangen: nämlich die genaue Form dieser Kurve zu modellieren, um so zu sehen, ob die Häufigkeiten in verschiedenen Stichproben (d.h. die spezifische Abnahme der Häufigkeiten) ein und dieselbe Form aufweisen oder nicht.

Dabei haben sich zwei Arten der Modellierung entwickelt: mit Hilfe von stetigen Funktionen oder diskreten Folgen und mit Hilfe von Wahrscheinlichkeitsfunktionen. Letztere unterscheiden sich von ersteren dadurch, dass sie normiert sind, d.h. dass die Summe der Wahrscheinlichkeiten 1 ergibt. Rein empirisch gesehen – bezüglich der Anpassungsgüte – haben reine Kurven einige Vorteile, theoretisch gesehen – nämlich bezüglich der Systematisierung – sind hingegen die Verteilungen adäquater.

Bei mehreren in der jüngsten Zeit durchgeführten Studien zu Graphemhäufigkeiten in slawischen Sprachen hat sich allerdings herausgestellt, dass in der Vergangenheit diskutierte und in der Gegenwart in der Regel immer noch in Betracht gezogene Modelle nicht passen, wenn man deren Anpassung an systematisch kontrolliertes Datenmaterial statistisch prüft. Demnach sind solche gängigen Modelle[4] wie etwa die geometrische Verteilung (Sigurd 1968) die Good-Verteilung (Martindale et al. 1996) und die Zipf'sche bzw. Zipf-Mandelbrot'sche Verteilung (vgl. Mandelbrot 1953) als Standardverteilung von Rangfrequenzen, die Waring-Verteilung oder die Whitworth-Verteilung (vgl. Whitworth 1901: 207f., Grzybek, Kelih, Altmann 2004) als passende Modelle oft nur lokal geeignet; die meisten nur in Kurvenform vorhandenen Modelle hingegen sind entweder nicht normiert, rechts nicht begrenzt oder ohne Interpretation der Parameter (vgl. Grzybek, Kelih, Altmann 2005a,b). Zumindest in den untersuchten slawischen Sprachen erweist sich als passendes Modell hingegen die negative hypergeometrische Verteilung (vgl. Grzybek, Kelih 2005; Grzybek, Kelih, Altmann 2004, 2005a,b).[5] Zwar handelt es sich bei dieser Verteilung um ein relativ komplexes Modell, dass nicht weniger als drei Parameter aufweist (*n, K, M*), doch haben sich im Zusammenhang mit den genannten Untersuchungen zunehmende Anhaltspunkte dafür ergeben, dass die Parameter *K* und *M* der negativen hypergeometrischen Verteilung sich letztendlich allein auf den Inven-

[4] Zur detaillierten Darstellung der einzelnen Verteilungsmodelle vgl. Wimmer/Altmann (1999).

[5] Auch für das Deutsche hat Best (2005a, 2005b) dieses Modell ins Spiel gebracht.

tarumfang *n* zurückführen lassen, wenn auch nicht auf direkte Art und Weise.

Es würde zu weit führen, diese Zusammenhänge hier im Detail zu diskutieren, weswegen auf die entsprechenden Arbeiten verwiesen sei. Allerdings haben sich aus dieser Beobachtung heraus weiterführende Überlegungen ergeben, die darauf abzielen, über das lokale Kriterium der in der Regel über den Chi²-Test geprüften Anpassungsgüte eines Modells hinausgehend auch globale Kriterien zu untersuchen, in erster Linie Kenngrößen wie die Entropie *H* oder die Wiederholungsrate *R*. Beide Kenngrößen sind asymptotisch ineinander überführbar (vgl. Altmann, Lehfeldt 1980), denn

$$H \approx ld\, n - \frac{nR-1}{2\ln(2)} \quad \text{und} \quad R \approx \frac{2(ld\, n - H)\ln(2)+1}{n}.$$

Versuchsweise sollte man aber zunächst beide ableiten. Der Sinn, diese globalen Kriterien zu bestimmen, liegt darin, dass man zusätzlich zum Kriterium der lokalen Eignung weiteres Adäquatheitskriterium erhält. Erweist sich ein globales Maß einer Verteilung als sehr allgemeingültig, dann kann man es als Modell auch dort verwenden, wo die Anpassungsgüte nicht besonders befriedigend ist. In solchen Fällen sucht man nach Gründen der Idiosynkrasie, z.B. nach Randbedingungen, wählt Spezialfälle oder Grenzfälle, modifiziert das grundlegende Modell oder sagt die weitere Entwicklung voraus.

Die von Herdan (1962: 36ff., 1966: 271ff.) in die Linguistik eingeführte Wiederholungsrate *R* [repeat rate, Herfindahlsches Konzentrationsmaß] ist eines der einfachsten globalen Charakteristiken einer Häufigkeitsverteilung, definiert als

$$(1) \qquad R = \sum_{k=1}^{n} p_k^2 .$$

Die Größe *R* bewegt sich im Intervall (1/*n*;1); sie erreicht den größten Wert, wenn ein einziges der vorkommenden Elemente die Wahrscheinlichkeit $p = 1$ hat und alle anderen Wahrscheinlichkeiten gleich Null sind; den kleinsten Wert nimmt *R* für den Fall an, dass alle Wahrscheinlichkeiten gleich sind. Daraus folgt, dass *R* ein Maß der Gleichverteilung ist, das desto kleiner wird, je ähnlicher die einzelnen Häufigkeiten sind. Auch die als

$$(2) \qquad H = -\sum_{i=1}^{n} p_i \cdot \mathrm{ld}\, p_i$$

definierte Entropie *H* lässt sich als ein Maß des Gleichgewichts bzw. der Gleichverteilung verstehen. Hier gilt, je größer H, desto ähnlicher die einzelnen Wahrscheinlichkeiten. Das Maximum (ld *n*) erreicht *H* dann, wenn alle Häufigkeiten gleich sind, das Minimum (0) hingegen, wenn eine der Wahrscheinlichkeiten $p_k = 1$. Da sich *H* somit im Intervall (0; ld *n*) bewegt, hängt sein Wert vom Inventarumfang ab, weswegen man zum Zwecke der Standardisierung auch die relative Entropie berechnet, die Werte im Intervall (0;1) annehmen kann.

$$(3) \qquad h = \frac{-\sum_{k=1}^{n} p_k \cdot \mathrm{ld}\, p_k}{\mathrm{ld}\, K}$$

Von diesen Überlegungen ausgehend, haben Altmann/Lehfeldt (1980) zunächst die (empirischen) Wiederholungsraten und Entropien für 63 Datensätze aus 38 verschiedenen Sprachen berechnet. Im Detail handelte es sich dabei um 26 Datensätze (von 23 verschiedenen

Sprachen), die auf Buchstaben- bzw. Graphemhäufigkeiten beruhten, und 37 Datensätze (von 27 verschiedenen Sprachen), die auf Phonemhäufigkeiten beruhten. Ohne zwischen Daten aus Buchstaben- bzw. Graphemhäufigkeiten und Daten aus Phonemhäufigkeiten zu differenzieren, haben die Autoren die erhaltenen Werte im Hinblick auf ein allgemeines Modell reflektiert. Da die vorliegende Abhandlung an diese Überlegungen anknüpft, bietet es sich an, die entsprechenden Daten sowohl in tabellarischer Form (vgl. Tab. 1) als auch in graphischer Form wiederzugeben; auf eine Darlegung der Quellen kann dabei im hier gegebenen Zusam-

Tabelle 1
Entropien und Wiederholungsraten von 63 Sprachen

Nr.	Sprache	*n*	*R*	*H*	*h*	Nr.	Sprache	*n*	*R*	*H*	*h*
1	Hawaiisch	13	0,120716	3,370804	0,910920	**33**	Ungarisch	32	0,053986	4,508421	0,901684
2	Hawaiisch B	13	0,131031	3,238512	0,875170	**34**	Ungarisch B	32	0,052090	4,544989	0,908998
3	Samoanisch	15	0,118269	3,475854	0,889673	**35**	Khasi	32	0,062487	4,468863	0,893773
4	Hawaiisch	18	0,105342	3,567112	0,855438	**36**	Lettisch B	32	0,056568	4,428232	0,885646
5	Pilipino	21	0,102547	3,820317	0,869773	**37**	Russisch B	32	0,056880	4,453237	0,890647
6	Pilipino	21	0,116018	3,725076	0,848089	**38**	Deutsch	33	0,059241	4,443530	0,880887
7	Kaiwa	21	0,080747	3,947756	0,898787	**39**	Georgisch	33	0,070258	4,293132	0,851070
8	See-Dajakisch B	21	0,091596	3,878690	0,883062	**40**	Georgisch	33	0,068390	4,310649	0,854542
9	Estnisch B	23	0,070229	4,086146	0,903033	**41**	Ostjakisch	33	0,062470	4,375786	0,867455
10	Suaheli B	24	0,087816	4,023958	0,877643	**42**	Ostjakisch	33	0,066969	4,395138	0,871292
11	Französisch B	24	0,079699	3,963404	0,864435	**43**	Ostjakisch	34	0,066856	4,406486	0,866146
12	Albanisch B	25	0,063719	4,217626	0,908216	**44**	Ostjakisch	34	0,064061	4,390940	0,863090
13	Indonesisch B	25	0,083864	3,928940	0,846051	**45**	Tschechisch	35	0,046491	4,700600	0,916424
14	Chamorro	25	0,074272	4,214724	0,907591	**46**	Tschechisch B	35	0,043964	4,722755	0,920744
15	Holländisch B	26	0,077732	4,085845	0,869247	**47**	Französisch	35	0,070591	4,101787	0,799680
16	Englisch B	26	0,062183	4,215092	0,896744	**48**	Marathi	38	0,060408	4,514403	0,860226
17	Rumänisch	27	0,062323	4,253963	0,894651	**49**	Bengali	38	0,065865	4,863256	0,926700
18	Spanisch B	27	0,074637	4,023919	0,846270	**50**	Ungarisch	39	0,052800	4,602810	0,870853
19	Hausa B	27	0,105588	3,922669	0,824976	**51**	Englisch	39	0,050495	4,709796	0,891095
20	Holländisch B	28	0,082022	4,048855	0,842221	**52**	Armenisch B	39	0,070748	4,433971	0,838909
21	Serbokroatisch B	29	0,063378	4,287102	0,882486	**53**	Russisch	41	0,050003	4,825688	0,900726
22	Bulgarisch B	29	0,067229	4,219317	0,868533	**54**	Polnisch	42	0,050928	4,725240	0,876291
23	Deutsch B	29	0,072007	4,172738	0,858945	**55**	Englisch	42	0,040990	4,918028	0,912044
24	Indonesisch	29	0,085824	4,094239	0,842786	**56**	Gujarati B	43	0,055151	4,664619	0,859637
25	Indonesisch	29	0,060762	4,348654	0,895157	**57**	Englisch	44	0,043749	4,906418	0,898705
26	Deutsch B	30	0,073938	4,147517	0,845243	**58**	Slovakisch	44	0,048530	4,731830	0,866726
27	Gujarati	30	0,055911	4,497795	0,916628	**59**	Schwedisch	45	0,043407	4,840576	0,881410
28	Italienisch B	30	0,074775	4,006747	0,816556	**60**	Ukrainisch	46	0,049402	4,627951	0,837856
29	Italienisch	31	0,067565	4,251224	0,858106	**61**	Hindi	52	0,054557	4,584604	0,804254
30	Ukrainisch B	31	0,049725	4,565024	0,921446	**62**	Burmanisch B	68	0,039244	5,230643	0,859248
31	Russisch B	31	0,057713	4,433604	0,894919	**63**	Vietnamesisch	74	0,047003	5,160549	0,831079
32	Amer. Englisch	32	0,054875	4,487366	0,897473						

menhang verzichtet werden (vgl. Altmann/Lehfeldt 1980: 154ff.).

Tab. 1 enthält die in der von Altmann/Lehfeldt (1980: 154ff.) angegebenen Reihenfolge Werte für die einzelnen Sprachen, wobei es sich bei den mit *B* bezeichneten Sprachen um Buchstaben- bzw. Graphemhäufigkeiten handelt. *K* bezeichnet die jeweilige Inventargröße, *R* ist der Wert für die Wiederholungsrate, *H* ist der Wert der absoluten, *h* derjenige der relativen Entropie. Man kann leicht sehen, dass sich die relative Entropie *h* in einem sehr schmalen Intervall von $0.7997 \leq h \leq 0.9267$ bewegt, was als ein weiteres Indiz für einen Zusammenhang mit dem Inventarumfang zu interpretieren ist.

Das Interesse von Altmann/Lehfeldt (1980) war es nun, an diese Daten ein allgemeines Modell anzupassen; dieses sollte nach Möglichkeit interpretierbar sein, womit im hier gegebenen Zusammenhang gemeint ist, dass das Modell letztendlich auf die Inventargröße zurückgeführt und in diesem Sinne erklärt werden kann. Zwar zogen die Autoren es auch in Betracht, zum Zwecke des Vergleichs ein Regressionsmodell des Typs $y = ax^{-b}$ anzupassen; doch ist bei diesem Modell keine Option erkennbar, wie die Parameter *a* und *b* auf den Inventarumfang projiziert und somit interpretiert werden könnten.

Deshalb haben Altmann/Lehfeldt – ausgehend von der geometrischen Verteilung – die sich aus dieser Verteilung ergebenden theoretische Entropien und Wiederholungsraten berechnet. Dabei waren die Autoren von der Idee geleitet, dass im Ergebnis ein geeignetes Modell zur Verfügung stehen würde, das theoretisch begründet wäre und über eine bestimmte Erklärungskraft verfügen würde.

Auf die mathematische Ableitung der theoretischen Entropie *E(H)* und Wiederholungsrate *E(R)* muss hier nicht im Detail eingegangen werden (vgl. Altmann/Lehfeldt 1980: 151ff.; Grzybek/Kelih/Altmann 2005c). Es mag ausreichend, dass sich die entsprechenden aus der geometrischen Verteilung ergebenden Werte für die Wiederholungsrate über die Formel

$$E(R) \approx \frac{2}{n} \tag{4}$$

berechnen und für die Entropie über die Formel

$$E(H) \approx -\mathrm{ld}\left[\left(\frac{4}{n+2}\right)\left(\frac{n-2}{n+2}\right)^{\frac{n-2}{4}}\right] \tag{5}$$

approximieren lassen.

Wie eine Re-Analyse der 63 Datensätze zeigt, stellt sich im Ergebnis für die Wiederholungsrate heraus, dass das Regressionsmodell gemäß der Formel $y = 0.9659x^{-0.7822}$ mit einen Determinationskoeffizienten von $R^2 = 0.7692$ eine bessere Anpassungsgüte auf weist als das sich aus der geometrischen Verteilung ergebende Modell ($R^2 = 0.6871$). In Abb. 1 sind die beobachteten Werte der Wiederholungsraten in Form von Datenpunkten markiert; enthalten sind auch die beiden Anpassungskurven: die durchgehende Kurve ist die Potenzkurve, die grob gestrichelte die sich aus der geometrischen Verteilung ergebende. Ebenfalls zu sehen ist eine weitere (fein gestrichelte) Linie. Hierbei handelt es sich um eine ebenfalls theoretisch begründete, sich aus der Zipf-Mandelbrot-Verteilung ergebende Kurve, die Zörnig/Altmann (1983) abgeleitet haben, da die sich aus der geometrischen Verteilung ergebenden theoretischen Werte der Wiederholungsrate zu keinen befriedigenden Ergebnissen geführt hatte. Wie die auf der entsprechenden Formel

$$(6) \qquad E(R) = \frac{1.61^{-1} - (0.61+n)^{-1}}{\left(\ln \frac{0.61+n}{1.61}\right)^2}$$

beruhende Re-Analyse der o.a. 63 Datensätze zeigt, sind die Ergebnisse mit einem Wert von $R^2 = 0.7427$ deutlich besser als die sich aus der geometrischen Verteilung ergebenden.

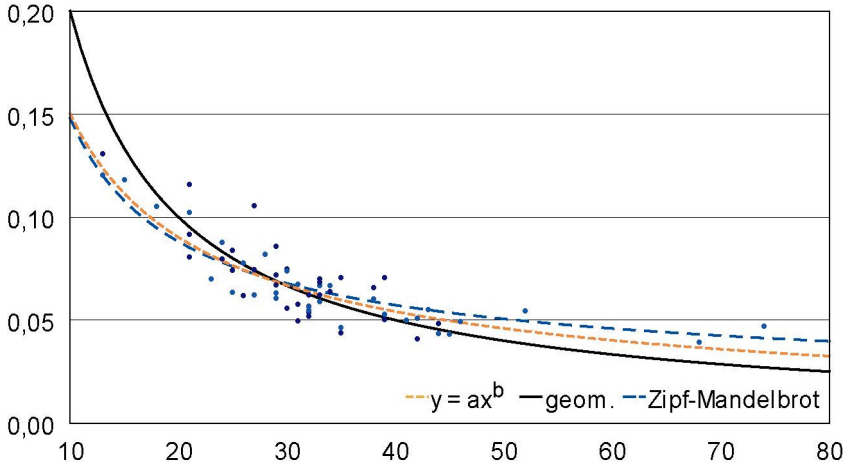

Abb. 1. Empirische und theoretische Wiederholungsraten für 63 Datensätze verschiedener Sprachen nach Altmann/Lehfeldt (1980)

Ein analoger Befund ergibt sich für die Berechnung der Entropie: Folgt man hier dem Modell $y = 1.7290x^{0.2663}$, so beträgt der Determinationskoeffizient $R^2 = 0.8538$; gemäß dem sich aus der geometrischen Verteilung ergebenden Modell (4) ist der Wert ebenso wie bei der Wiederholungsrate bei $R^2 = 0.7928$ deutlich schlechter. Abb. 2 stellt den Befund anschaulich dar, die Form der Darstellung entspricht der in Abb. 1. Auch Abb. 2 enthält zusätzlich die sich aus der Zipf-Mandelbrot-Verteilung ergebenden theoretischen Entropiewerte, die nach der von Zörnig/Altmann (1984) abgeleiteten Formel

$$(7) \qquad E(H) = \mathrm{ld}\, e \ln\left[\sqrt{(B+n)(B+1)}\,\ln\frac{B+n}{B+1}\right]$$

mit $B = 0.61$ zu einem Wert von $R^2 = 0.8371$ führt, der ebenfalls besser ist als der sich aus der geometrischen Verteilung ergebende.

In beiden Fällen ist das theoretisch begründbare Modell im Vergleich zu dem rein rechnerisch optimierten somit das – relativ gesehen – schlechtere. Doch hat das aus der Potenzformel gewonnene Modell den entscheidenden Nachteil, dass es rein empirisch aus den konkreten Daten gewonnen ist und nicht interpretierbar. Abgesehen davon, dass es bei hinzukommenden Daten folglich jeweils (mehr oder weniger erheblich) modifiziert werden müsste, ist es schlicht und einfach nicht interpretierbar. Die aus der geometrischen Verteilung bzw. der Zipf-Mandelbrot-Verteilung gewonnenen Kurven hätten hingegen den Vorteil einer Erklärungs- bzw. Prädiktionsoption, die vor allen Dingen darin bestünde, dass die globalen Masse (Entropie und Wiederholungsrate) ausschließlich als von der Inventargröße abhängig interpretiert werden könnten. Vor diesem Hintergrund wird es von Interesse sein, die Ergebnisse mit den sich aus den anderen eingangs erwähnten Verteilungen (Good-Verteilung, Whitworth-

Verteilung, negative hypergeometrische Verteilung) zu vergleichen (Grzybek, Kelih, Altmann 2005c).

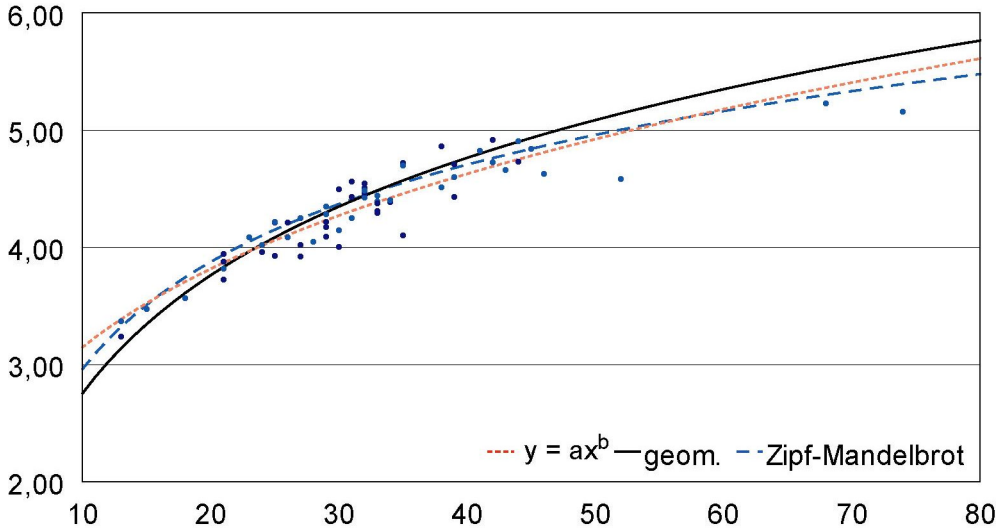

Abb. 2. Empirische und theoretische Entropien für 63 Datensätze verschiedener Sprachen nach Altmann/Lehfeldt (1980)

Hier aber soll es um eine andere Fragestellung gehen. Denn die bislang angesprochenen Studien haben mitsamt eine entscheidende Frage offen gelassen: es wurde in den genannten Studien nämlich nicht zwischen der Häufigkeit von Graphemen und Phonemen unterschieden, in der Annahme, dass beide Einheiten ohnehin ein und denselben Regularitäten folgen. Ohne Frage ist dies eine plausible Annahme, dennoch ist diese bislang ungeprüft. Und genau an dieser Stelle setzen die hier vorliegenden Überlegungen ein, die im vorliegenden Text aufgegriffen werden sollen: Denn wenn sich – wie eingangs gesagt wurde – in der konkreten empirischen Untersuchung zumindest mehrerer slawischer Sprachen gezeigt hat, dass weder die geometrische Verteilung noch die Zipf-Mandelbrot-Verteilung geeignete Modelle für Ranghäufigkeiten von Graphemen oder Phonemen darstellen, dann ergibt sich zwangsläufig die Frage, ob sich die theoretischen Werte für Entropie und Wiederholungsrate überhaupt an einer dieser Verteilungen orientieren sollten.

Bevor das aber konkret geprüft wird, gilt es im hier vorliegenden Zusammenhang der erwähnten impliziten Annahme nachzugehen, nämlich der Tatsache, dass die Datensätze der 63 Sprachen auf heterogenem Material basieren: Damit ist nicht einmal gemeint, dass den Erhebungen der Häufigkeiten unterschiedliche Graphem- und/oder Phonem-Definitionen für eine gegebene Sprache oder zwischen den Sprachen zugrunde liegen, dass in einigen Untersuchungen Satz- und Leerzeichen als eigenständige Elemente berücksichtigt wurden und in anderen nicht, usw. Gemeint ist vielmehr, dass Phoneme und Grapheme undifferenziert in gleicher Art und Weise behandelt wurden, obwohl es sich – linguistisch gesehen – um unterschiedliche Arten der Repräsentation und auch Abstraktion handelt. Altmann/Lehfeldt (1980) sind bei diesem Vorgehen der Annahme gefolgt, dass in beiden Fällen[6] die theoretischen Modelle letztendlich auf einen einzigen Faktor, nämlich den des Inventarumfangs, zurück-

[6] Auch wenn sich weitere Differenzierungen etwa zwischen Buchstaben und Graphemen, oder zwischen Lauten und Phonemen, berücksichtigen ließen, soll hier lediglich von „zwei Fällen“ der Repräsentation – nämlich Graphemen und Phonemen – die Rede sei, zumal sich derartige weitere Differenzierungen nicht ohne erheblichen Aufwand aus dem verwendeten Material rekonstruieren ließen.

zuführen sind. Dennoch aber fehlt bis heute der Nachweis, dass sich das Verhalten der Phoneme und der Grapheme in dieser Hinsicht nicht nachhaltig voneinander unterscheidet – und bevor weitere, aus anderen Verteilungen hervorgehende theoretische Entropien und Wiederholungsraten erarbeitet werden, scheint es angebracht, diesen Umstand empirisch zu testen und gegebenenfalls die von Altmann/Lehfeldt (1980) implizit vertretene Ansicht zu festigen.

Diese Überprüfung kann freilich sinnvollerweise nur an den empirischen Daten vorgenommen werden, woraus sich eine mehrstufige Re-Analyse der von den Autoren verwendeten Daten und eine multiple Berechnung der Entropien und Wiederholungsraten ergibt:

1. Anpassung des nicht-linearen Regressionsmodells nach der Formel $y = ax^{-b}$
   a. für das gesamte, nicht nach Graphemen und Phonemen differenzierte Datenmaterial;
   b. für die auf Graphemen basierenden Datensätze;
   c. für die auf Phonemen basierenden Datensätze;

2. Vergleich der Regressionskurven

Vor dem Hintergrund der obigen Ausführungen mag es verwunderlich erscheinen, dass der angestrebte Vergleich auf der Basis der Potenzkurve vorgenommen werden soll, die ja – wie argumentiert wurde – theoretisch nicht begründbar ist. Solange wir aber nicht wissen, welcher theoretische Ansatz am adäquatesten ist – d.h. welcher die meisten Kriterien, die wir in einer anderen Arbeit erörtern werden, erfüllt – beschränken wir uns auf eine empirisch gut passende Kurve, die gleichzeitig eine Maßlatte für andere Kurven darstellt. Findet man eine theoretische Kurve, die gleich gut (bzw. nicht viel schlechter) ist wie die empirische, dann wird man natürlich immer die theoretisch begründete vorziehen.

Wiederholungsraten

Abb. 3a zeigt die Wiederholungsraten für die 63 Datensätze, mit unterschiedlichen Markierungen für graphem- und phonembasierte Daten. Abb. 3b zeigt die Anpassungslinie gemäß der Formel $y = ax^{-b}$. Bei Werten für $a = 0.817766$ und $b = -0.735897$ beträgt die Güte der Anpassung $R^2 = 0.7177$.

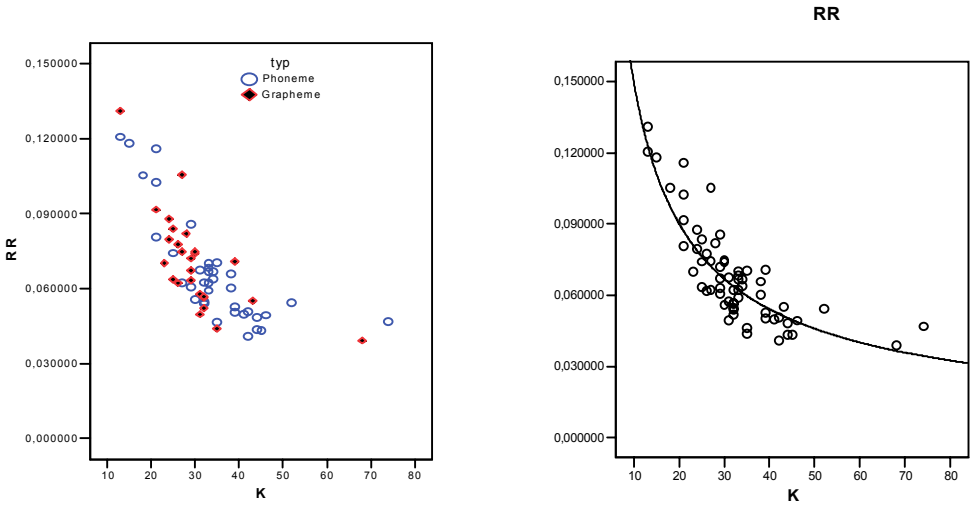

Abb. 3a/3b. Daten und Anpassung für Grapheme und Phoneme

Abb. 4a/b zeigt die Anpassungsergebnisse für die nach Graphemen und Phonemen differenzierten Datensätze: Für die Datensätze der Grapheme beträgt die Anpassungsgüte $R^2 = 0.6348$ bei Werten von $a = 0.8369$ und $b = –0.7449$, für die Datensätze der Phoneme erhält man bei Werten von $a = 0.8243$ und $b = –0.7368$ eine Anpassungsgüte von $R^2 = 0.7593$.

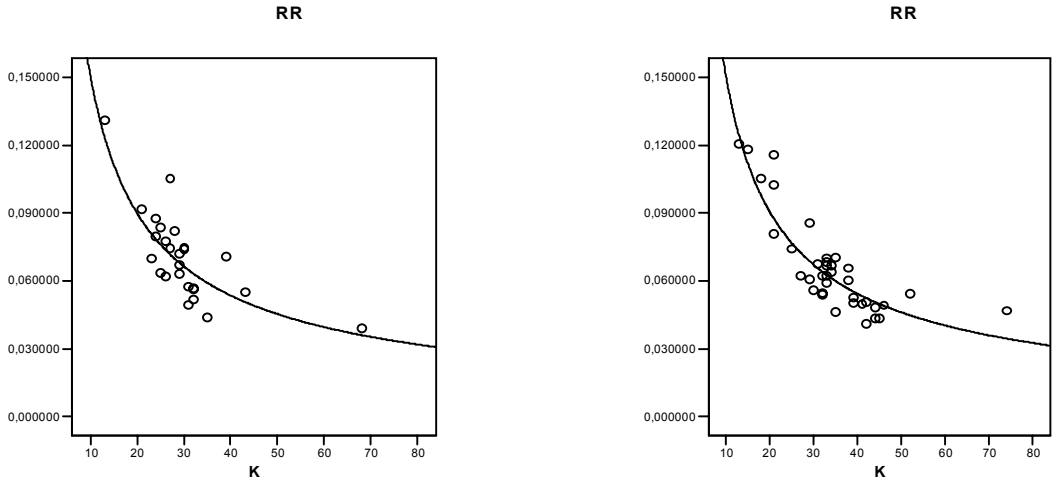

Abb. 4a/4b. Getrennte Anpassung der Formel $y = ax^{-b}$ an die Datensätze der Grapheme (links) und der Phoneme (rechts)

Auch wenn auf den ersten Blick die Kurvenverläufe sehr ähnlich wirken, ist es interessant und notwendig, die beiden Regressionskoeffizienten miteinander zu vergleichen und statistisch auf Unterschiede zu testen. Dies geschieht bei linearen Zusammenhängen über die *t*-verteilte Prüfungsgröße

$$t = \frac{|b_1 - b_2|}{\sqrt{\frac{s^2_{y1.x1} \cdot (n_1 - 2) + s^2_{y2.x2} \cdot (n_2 - 2)}{n_1 + n_2 - 4} \cdot \left(\frac{1}{Q_{x1}} + \frac{1}{Q_{x2}} \right)}}$$

bei $FG = n_1 + n_2 - 4$ Freiheitsgraden mit $Q_x = \sum (x - \overline{x})^2$; s^2 sind die üblichen Varianzen.

Um also die beiden Regressionskoeffizienten der Graphem- und der Phonemdaten auf Unterschied zu testen, ist eine einfache Linearisierung der Gleichung $y = a \cdot x^b$ notwendig, die sich durch eine einfache Logarithmierung in $\ln(y) = \ln(a) + b \ln(x)$ umformen lässt.

Im Ergebnis stellt sich heraus, dass der Unterschied zwischen den beiden Regressionskoeffizienten bei einem Wert von $t_{FG=59} = 0.0622$ nicht im mindesten signifikant ist ($p = 0.95$).

Entropie

Abb. 5a zeigt die Entropien für die 63 Datensätze, mit unterschiedlichen Markierungen für graphem- und phonembasierte Daten. Abb. 5b zeigt die Anpassungslinie gemäß der Formel $y = ax^{-b}$. Bei Werten für $a = 1.6602$ und $b = –0.2779$ beträgt die Güte der Anpassung $R^2 = 0.8672$.

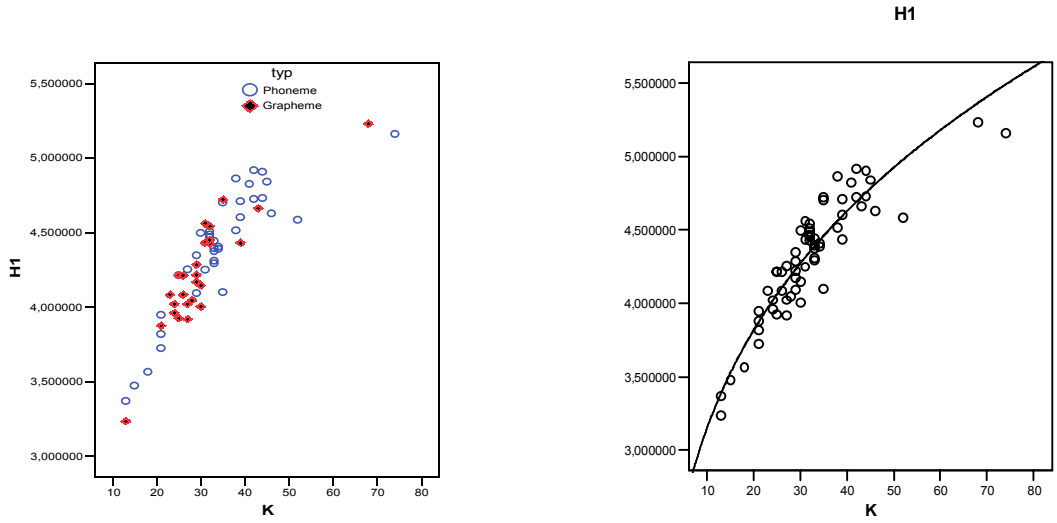

Abb. 5a/5b. Daten und Anpassung für Grapheme und Phoneme

Abb. 6a/b zeigt die Anpassungsergebnisse für die nach Graphemen und Phonemen differenzierten Datensätze: Für die Datensätze der Grapheme beträgt die Anpassungsgüte $R^2 = 0.8514$ bei Werten von $a = 1.6037$ und $b = 0.2875$, für die Datensätze der Phoneme erhält man bei Werten von $a = 1.7002$ und $b = 0.2715$ eine Anpassungsgüte von $R^2 = 0.8697$.

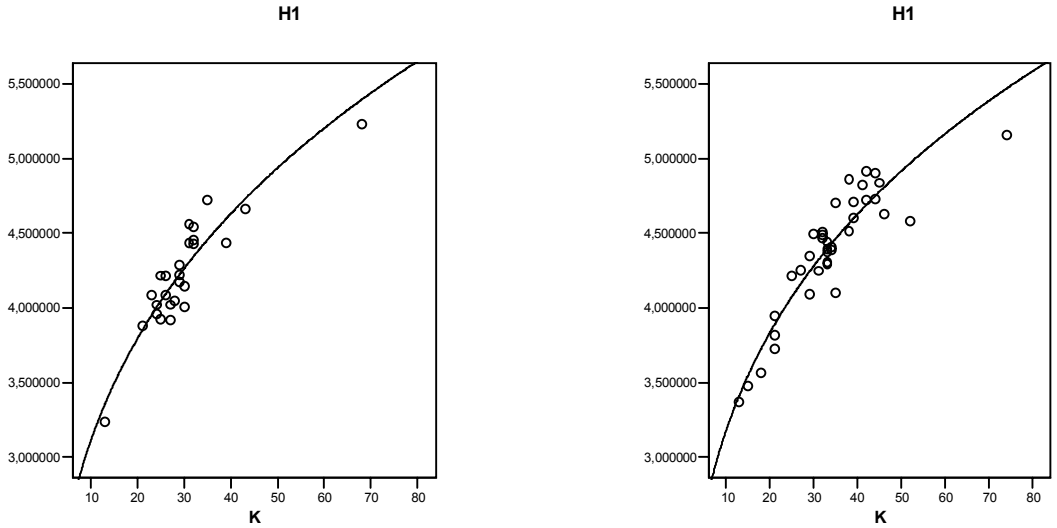

Abb. 6a/6b. Getrennte Anpassung der Formel $y = ax^{-b}$ an die Datensätze der Grapheme (links) und der Phoneme (rechts)

Auch hier stellt sich im Ergebnis heraus, dass der Unterschied zwischen den beiden Regressionskoeffizienten bei einem Wert von $t_{FG=59} = 0.5229$ nicht signifikant ist ($p = 0.60$).

Zusammenfassung

Eine wesentliche Schlussfolgerung aus den obigen Überlegungen und Untersuchungen ist es, dass Phoneme, Buchstaben und Grapheme sich im Hinblick auf ihre Häufigkeitsverteilung

gleich verhalten; das ist insofern nicht unbedingt erstaunlich, da die betreffenden Systeme im Prinzip die gleiche Funktion erfüllen. Damit soll natürlich nicht gesagt sein, dass es unerheblich ist, welche dieser Einheiten in einer gegebenen Sprache bzw. für eine gegebene Sprache erhoben und untersucht werden – vielmehr sollte grundsätzlich die Qualität der untersuchten Einheiten sauber differenziert werden.[7] Mit gleichem Verhalten ist gemeint, dass sich die theoretischen Modelle zur Beschreibung von Häufigkeiten all dieser Einheiten letztendlich ausschließlich auf den Inventarumfang *n* zurückführen lassen und somit auch interpretieren lassen sollten. Dafür jedenfalls sprechen die oben angeführten Analysen und Re-Analysen.

Literatur

Altmann, G. (1993). Phoneme counts. *Glottometrika 14, 54-58.*

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

Best, K.-H. (2005a). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu.* [Im Druck]

Best, K.-H. (2005b). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11* (eingereicht).

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. *Anzeiger für Slavische Philologie 31, 131-162.*

Grzybek, P., Kelih, E. (2005). Grapheme frequencies in Slovene. In: Benko, Vladimir (ed.), *Slovko 2003*. Bratislava. [Im Druck].

Grzybek, P., Kelih, E., Altmann, G. (2004). Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie 32, 25-54.*

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

Grzybek, P., Kelih, E., Altmann, G. (2005b). Graphemhäufigkeiten im Slowakischen (Teil II: Mit Digraphen). In: *Sprachen und Sprache im Mitteleuropäischen Raum*. [Im Druck]

Grzybek, P., Kelih, E., Altmann, G. (2005c). Grapheme frequencies: Model characteristics and criteria. In: *Journal of Quantitative Linguistics.* [In Vorb.]

Herdan, **G.** (1962). *The calculus of linguistic observations.* The Hague: Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance.* Berlin: Springer.

Hussien, *O.A*. (2004). The Lerchianness plot. *Glottometrics 7,* 50-64.

Mandelbrot, **B. B.** (1953). An information theory of the statistical structure of language. In: W. Jackson (ed.), *Communication Theory: 503-512*. New York: Academic Press.

Martindale, C., McKenzie, D., Gusein-Zade, S.M., Borodovsky, M.Y. (1996). Comparison of equations describing the frequency distribution of graphemes and phonemes. *J. of Quantitative Linguistics 3, 106-112.*

---

[7] So sollte etwa zwischen Buchstaben, Graphemen und Phonemen sowie deren jeweiligen Häufigkeiten deutlich getrennt werden – wobei unter Buchstaben Einzelzeichen und unter Graphemen auch (!) Kombinationen von Buchstaben zu verstehen wären, so z.B. [sch] im Deutschen, [sh] im Englischen, [ch] im Französischen, [cs, sz, ny, gy,…] im Ungarischen usw. Vor diesem Hintergrund wäre davon auszugehen, dass innerhalb einer gegebenen Sprache für die Inventargröße der drei Einheiten die ungefähre Regel gilt:

Buchstabeninventar ≤ Grapheminventar ≤ Phoneminventar,

wobei es zahlreiche Ausnahmen gibt.

Naranan, S., Balasubrahmanyan, V.K. (1992a). Information theoretic models in statistical linguistics. Part I: A model for word frequencies. *Current Science 63, 261-269.*

Naranan, S., Balasubrahmanyan, V.K. (1992b) Information theoretic models in statistical linguistics. Part II: Word frequencies and hierarchical structure in language – statistical tests. *Current Science 63, 297-306.*

Orlov, Ju.V., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen*. Bochum: Brockmeyer.

Pääkkönen, M. (1993). Graphemes and context. *Glottometrika 14, 1-53*.

Rosenbaum, R., **Fleischmann, M.** (2002). Character frequency in Multilingual Corpus 1 – Part 1. *Journal of Quantitative Linguistics 9(*3), *233-260*.

Rosenbaum, R., **Fleischmann, M.** (2003). Character frequency in Multilingual Corpus 1 – Part 2. *Journal of Quantitative Linguistics 10(1)*, *1-39*.

Sigurd, B. (1968). Rank-frequency distribution for phonemes. *Phonetica 18, 1-15.*

Tuldava, J. (1988). Opyt kvantitativnogo analiza sistemy fonem ėstonskogo jazyka*. Acta et Commentationes Universitatis Tartuensis 838, 120-133.*

Whitworth, **W.A.** (1901). *Choice and chance. With one thousand exercises*. New York, London 1965.

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

Zörnig, P., Altmann, G. (1983). The repeat rate of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 5, 205-211*.

Zörnig, P., Altmann, G. (1984). The entropy of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 6, 41-47.*

Glottometrics 9

2005

RAM - Verlag

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –9 (2005) –. – Lüdenscheid: RAM-Verl., 2005
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 9 (2005)
ISSN 1617-8351