Glottometrics 15, 2007, 82-91

On the systematic and system-based study of grapheme frequencies:
a re-analysis of German letter frequencies

Peter Grzybek, Graz

Abstract: This study looks at the theoretical modeling of letter frequencies. Based on recent findings demonstrating the negative hypergeometric function to be an adequate model, a re-analysis of German data reported by Best (2005) is conducted, concentrating on a detailed examination of parameter behavior. It is shown that all parameters of this distribution behave regularly, if the analysis is based on the system's inventory size, rather than on the class of items occurring in the given sample. Directions for future research are pointed out, particularly involving factors influencing parameter values.

Keywords: Grapheme frequency, German, negative hypergeometric distribution

Introduction

The frequency of letters and graphemes has recently been the focus of a growing level of interest. This holds particularly true with regard to various Slavic languages, which, in the last years, have been submitted to systematic studies, starting with Russian (Grzybek, Kelih 2003; Grzybek 2005; Grzybek, Kelih, Altmann 2004, 2005a), and including Slovak (Grzybek, Kelih, Altmann 2005b, 2007), Ukrainian (Grzybek, Kelih 2005b), Slovene (Grzybek, Kelih, Stadlober 2007).[1] These Slavic languages cover a spectrum of grapheme inventory size, from the minimal inventory of 25 letters (Slovene) to the maximum of 43 or 46 letters (Slovak)[2].

With regard to other languages, or language families, similarly systematic studies are not available, neither with regard to the material analyzed nor with regard to the theoretical models aiming to describe the frequencies and their distributions. Only for German grapheme frequencies are comparable studies available, from Karl-Heinz Best's (2005) study searching for regularities in the frequency behavior of letters and other characters. Best (2005: 9) starts from the assumption that, for German, there are only relatively sparse data which, furthermore, are quite obsolete and therefore give rise to the question of whether "they are still representative for contemporary circumstances". Since, additionally, these data are based on the analysis of heterogeneous corpus material, rather than on individual texts, Best (2005: 11) has pursued the question of whether "there is a theoretically motivated model which might be adequate to represent empirical data from rank frequency distributions of letters and other characters".

In this text, Best's specific data shall be submitted to an elaborating re-analysis. Before going into details as to Best's data, it seems reasonable, however, to briefly summarize the general framework of the overall problem.

[1] For each of these languages, series of 30 samples have been systematically analyzed, partly controlling authorship, text type, and homogeneity of texts (text segments, text cumulations, text mixtures, etc.) as possible influencing factors.

[2] The difference in the Slovak inventory size depends on whether the three digraphs DZ, DŽ, CH are treated as separate inventory units or not; Grzybek, Kelih & Altmann (2005b) and Grzybek, Kelih & Altmann (2007) have studied both alternatives separately.

1. The negative hypergeometric distribution as a rank frequency model

Studies of rank frequencies focus on the proportion of the most frequent unit as compared to the second, third, etc. one, that is, on the overall relation between the individual frequencies. The objective of this approach is the theoretical modeling of such a rank frequency distribution, searching for a mathematical formalization of the distances between the individual occurrences: transforming the initial raw data into a ranked (usually decreasing) order, and connecting the individual data points, usually, a particular declining (hyperbolic) curve is obtained, rather than a linear decrease. It is the mathematical modeling of this curve which is at the center of this field of research, to see whether or not if the frequencies (or rather, the shape of their specific decline) is similar across different samples.

At closer sight, graphemes and their rank frequency distributions represent a discrete system, not a continuous curve. Since we are therefore rather concerned with two neighboring classes, it seems reasonable to search for an adequate discrete probability distribution, rather than for a continuous function; this has other advantages, too, which need not be mentioned in detail here (see Grzybek, Kelih 2003). In this context, referring to the theoretical framework of synergetic linguistics, we are faced with the generally accepted assumption that the probability of a given class *x* (or rank *r*) behaves itself proportionally to the neighboring lower class, i.e., *x*-1, or *r*-1 (see Altmann, Köhler 1996). Based on this general approach we formulate the difference equation

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

the concrete solution of which depends on the concrete form of the function *g(x)*. As to the frequency of various linguistic units, relatively simple functions have repeatedly been shown to yield convincing results, even with *g*(*x*) being represented by simple rational functions. In attempting to qualitatively interpret these functions, the “speaker’s forces” were assumed to be represented in the function’s numerator, the regulating “hearer’s forces”, as compared to this, in its denominator. This approach has recently been significantly generalized by Wimmer, Altmann (2005, 2006); for linguistic questions, various distribution models, among others, can be derived from the central equation:

$$(2) \qquad P_x = \left(1 + a_0 + \frac{a_1}{(x+b_1)^{c_1}} + \frac{a_2}{(x+b_2)^{c_2}}\right)P_{x-1}$$

One of these models is the negative hypergeometric distribution (Wimmer, Altmann 1999: 465ff.), which, in all above-mentioned studies, has turned out to be an adequate model for both the Slavic languages and German. After re-parametrization, from (2) the recursion formula (3) is obtained

$$(3) \qquad P_x = \frac{(M+x-1)(n-x+1)}{x\left(K-M+n-x\right)}P_{x-1} \ ,$$

from which the negative hypergeometric distribution results:

$$(4) \qquad P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}} \qquad x = 0,1,2,\ldots,n, \quad K > M > 0;\ n \in \{1,2,\ldots\}$$

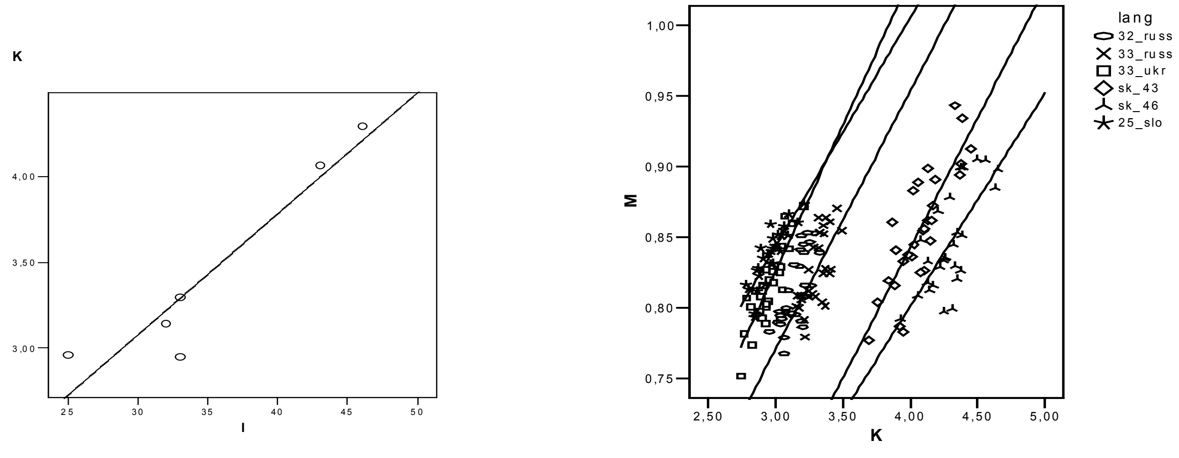
For ranking purposes, this distribution is conventionally shifted one step to the right, thus yielding the 1-displaced negative hypergeometric distribution (5):

$$(5)\qquad P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}} \qquad \begin{array}{l} x = 1, 2, \ldots, n+1 \\ \\ K > M > 0;\ n \in \\ \{1,2,\ldots\} \end{array}$$

2. From model to parameter interpretation

The adequacy of the distribution model (5) for letter frequencies has been repeatedly demonstrated in the above-mentioned recent research, for Slavic languages as well as for German. In the case of the Slavic languages, the adequacy of other models previously discussed in linguistics has been also tested in a systematic way. Only the negative hypergeometric distribution has turned out to be an overall valid model. Furthermore, the analyses of Slavic letter frequencies have yielded the first insight into the systematic behavior of parameters *K* and *M*, thus allowing for some hypotheses as to their qualitative interpretation (on this question, see Grzybek, Kelih 2005c; Grzybek et al. 2006). Yet attempts at general parameter interpretation are in their infancy; to achieve this goal, further studies, examining more languages, are needed. From what we know, it seems that three factors have particular impact on the parameter values: 1) inventory size, 2) relative frequency of the first rank, and 3) mean value of the given distribution.

It seems that inventory size is the foremost influence on the overall system behavior. Figures 1a and 1b illustrate the general tendency as it emerges from the analysis of six sample series from Slovene (*I* = 25), Russian (*I* = 32, or *I* = 33, respectively)[3], Ukrainian (*I* = 33) and Slovak (*I* = 43, or *I* = 46, respectively)[4]: Figure 1a shows the dependency of parameter *K* on inventory size *I*, based on the parameter mean values of each language; with a correlation coefficient of $r = 0.94$ the linear dependence turns out to be significant ($p = 0.005$). Figure 1b shows the correlation between parameters *K* and *M*, which is not, however, relevant between languages, but rather within a given language.



(a) Correlation between *K* and *I* (b) Correlation between *K* and *M*

Figure 1. Parameter Behavior for six Slavic Sample Series

[3] The difference in the Russian inventory size depends on treating the letter ‚ë' as a separate letter in its own right or not (cf. Grzybek et al. 2005).

[4] As to the differences in Slovak inventory size, see fn. 1.

As shown by the graphs, the regression lines for the individual languages display a clear tendency to be parallel, which, in turn, can also be interpreted in terms of a dependency on inventory size. The regression lines follow the equation $y = b + ax$ (that is, in our case, $M = b + aK$); here, b is a constant determining the regression intercept, and a is the regression coefficient which determines the steepness for the rise or decline of the line. Introducing the intercept values of the individual languages into a regression model with inventory size I as the independent variable yields a highly significant correlation ($r = 0.96$, $p < 0.001$); Figure 2 illustrates this correlation.

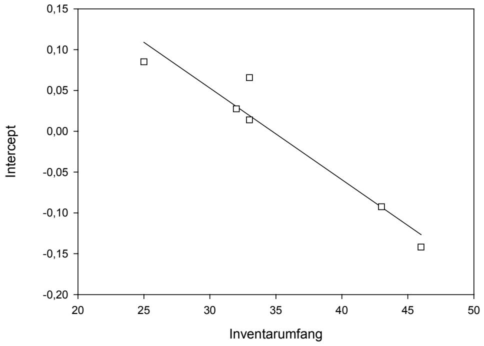


Figure 2. Correlation between the intercepts of the regression model and inventory size *I* for six Slavic sample series

So we are concerned, on the one hand, with an interlingual (linear) dependence of parameter K on inventory size I; and on the other hand, with a language-specific (linear) dependence of parameter M on K. For all the languages studied previously (Russian, Slovak, Slovene, Ukrainian), this situation can be traced back to an overall (linear) regression model, for which interlingual and language-specific principles can be distinguished (Grzybek et al. 2006).

The study of other languages has not yet reached this point; this holds also true for German, where the question of parameter interpretation has not yet been touched upon. Given this state of the art, the following re-analysis of the data presented by Best (2005) focuses on the systematic study of the parameter values obtained for the negative hypergeometric distribution.

2. Material

As mentioned above, Best (2005) provides a number of analyses of individual texts, in addition to corpus data; the resulting 14 data sets are characterized in Table 1:

Table 1
Text basis from Best (2005)

Nr.	Text	Nr.	Text
1	H. Pestalozzi: Hühner, Adler und Mäuse	**8**	F. Kafka: Der Prozeß
2	G.A. Bürger: Münchhausen	**9**	G. Vesper: Fugen
3	G.A. Bürger: Lenore	**10**	O. Jägersberg: Dazugehören
4	G. Büchner: Lenz	**11**	J. Joffe: Nach dem Bruderkrieg
5	G. Büchner: Hessischer Landbote	**12**	R. Hoppe: Das gierige Gehirn
6	K. May: Winnetou I	**13**	Schönpflug (1969)
7	F. Kafka: Die Verwandlung	**14**	K.H. Best: Wiss. Prosa

In Best's (2005) study, some of these texts have been analyzed twice: once taking into account "only" letters, and once including all occurring characters (such as blanks, apostrophes, dashes, etc.). Therefore, Best's study contains not only 14, but rather19 analyses. It goes without saying that taking into account these additional characters not only changes individual letters' relative frequency, but also the inventory size. Table 2 presents the relevant characteristics of the data: the column "Table" refers to Best's original numeration of tables, "Data Set" to the corresponding data set(s), partly analyzed twice. *N* indicates samples sizes, *I* is the corresponding inventory size. Table 2 also contains the values of parameter values *K* and *M* of the negative hypergeometric distribution, as given by Best (2005), as well as the corresponding fitting results, *C* being the determination coefficient calculated as X^2 / *N*.[5]

Table 2
Data sets from Best (2005)

Table	/ Data Set	*N*	*I*	*K*	*M*	X^2	*C*
1	1	675	27	3,0071	0,6847	17,63	0,0261
2	2	137476	30	3,4096	0,7385	1311,63	0,0095
3	3a	6215	27	3,1886	0,8109	45,99	0,0074
4	3b	7962	37	3,0934	0,6574	51,71	0,0065
5	4a	42608	30	3,4083	0,7289	425,28	0,0100
6	4b	53443	47	4,8953	0,6991	750,3	0,0140
7	5	21452	30	3,3167	0,7016	147,25	0,0069
8	6a	777368	32	3,7659	0,7610	9973,4	0,0128
9	6b	974506	48	4,7707	0,7033	12316,23	0,0126
10	7	99559	30	3,2877	0,7218	954,41	0,0096
11	8	361848	30	3,3554	0,7350	3437,86	0,0095
12	9a	6259	27	2,9541	0,7251	46,28	0,0074
13	9b	7555	31	3,0799	0,6859	92,00	0,0122
14	10	40977	30	3,4964	0,7775	269,79	0,0066
15	11	6091	30	3,4038	0,7372	35,42	0,0058
16	12a	20075	30	3,3278	0,7366	103,44	0,0052
17	12b	24977	53	6,4117	0,7414	305,04	0,0122
18	13	99984	29	3,2200	0,7265	462,28	0,0046
19	14	179922	30	3,2002	0,7254	1423,51	0,0079

4. Results

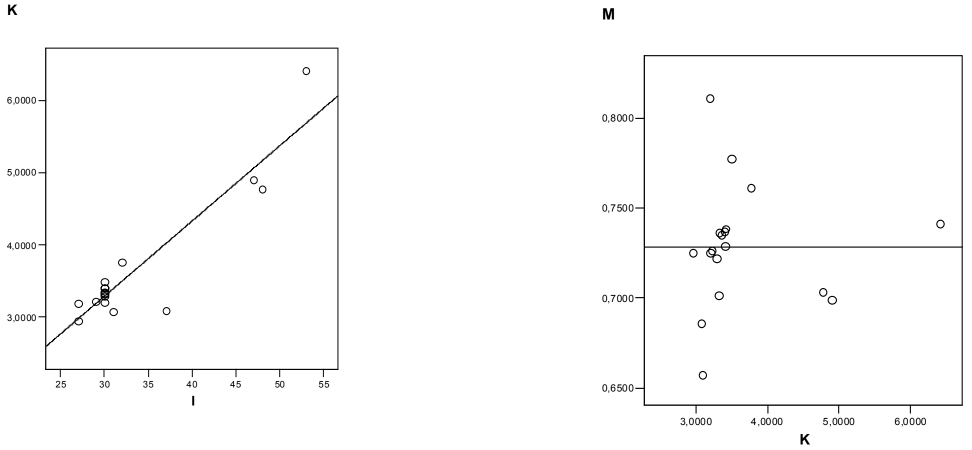
Without exception, the negative hypergeometric distribution turns out to be a very good model: With the exception of the very first text (extremely short at 675 letters), the values of the discrepancy coefficients are in the interval of $0.0046 \leq C \leq 0.0140$ – thus proving the negative hypergeometric distribution indeed to be a good model.[6]

[5] Interpreting the goodness of fit with reference to the X² values would have to be based on D*F* = *I*-4 degrees of freedom. Since the X^2 value increases linearly with sample size, and in this case tends to yield significant results more quickly, linguistic studies concerned with large sample sizes rather use to refer to the discrepancy coefficient. By way of convention, a value of $C \leq 0.02$ is interpreted as indicating a good, a value of $C \leq 0.01$ a very good fit; the degrees of freedom are irrelevant, here.

[6] The fit is very good, for the first text, too, with a X^2 value of 17.63, with 22 degrees if freedom corresponding to $P = 0.73$; the extreme differences in sample sizes, however, ranging from 675 to 974506, allows for a comparison of the longer texts only; therefore text #1 is excluded from the following re-analyses.

As mentioned above, due to Best's specific design partly including not only letters, the inventory sizes vary significantly in the interval of $27 \leq I \leq 53$. In analogy to the tendency described above for Slavic languages, this corresponds to a relatively large range for parameters *K* and *M*, which are in the intervals of $2.95 \leq K \leq 6.41$, and $0.66 \leq M \leq 0.81$ respectively.[7] Table 2 presents the results in detail.

Figures 3a and 3b illustrate the relations between inventory size and parameter *K*, and between parameters *K* and *M*.



(a) Relation between *K* and *I* (b) Relation between *K* and *M*

Figure 3. Parameter Behavior for 18 German Samples (Data reported by Best 2005)

The significant ($r = 0.92$) correlation between *K* and *I* can clearly be seen; no specific relation can be detected, however, between parameters *K* and *M*. As opposed to the studies reported above that concentrated on Slavic languages, the results are thus far from easily interpretable, particularly because the parameter values obtained for *K* and *M* display no systematic behavior. As an explanation for this lack of systematic behavior, it seems reasonable to assume that it is due to the varying inventory size, as a consequence of the changing number of units submitted to analysis.

In consequence, for the sake of a consistent and unitary treatment of the data material reported by Best (2005), the latter shall be submitted to a comprehensive re-analysis concentrating on the analysis of letters, only, and excluding all other characters.

At closer examination, however, a problem comes into play, which is not discussed in Best's study and, as a consequence, not treated systematically. This problem concerns the unitary treatment of letters which do not occur in a given sample. Basically assuming an inventory size of $I = 30$ for German[8], Best (2005) confines the inventory size to $I = 27$ for those cases – such as, for instance, #1, #3, or #9 – where letters Q,X,Y do not occur. Similarly, sample #13 – where letter ß does not occur – is restricted to (and calculated as) inventory size $I = 29$. In other cases – e.g., #7 und #10, where there is no X or no Y – Best (2005) has allocated frequency $f_i = 0$ to these classes; as a consequence, the inventory size of these samples is $I = 30$, notwithstanding the fact that in the corresponding case, these letters are missing from the data material. Furthermore, Best (2005) assumes sample #6 has an inventory size of $I = 32$, given the fact that a number of foreign words with the letters É, Ñ occur in the material.

[7] Upper and lower borders of the 95% confidence interval vary significantly for parameter *K* as well (3.21 and 4.08); as compared to this, the confidence interval for *M* is much small with upper and lower borders of 0.71 and 0.75, respectively.

[8] This definition pays no attention to the distinction between lowercase and capital letters, considering Ä, Ö, Ü, ß as separate units in their own right.

As a consequence, parameter *n* of the negative hypergeometric distribution ranges from 26 (e.g., when letters Q, X, Y do not occur in a given sample), to 28 (when ß is not taken into consideration), to 29, in one case even to 31, since in this text (*Winnetou I* by Karl May) É and Ñ happen to occur and are considered to be elements of the inventory. Table 2 represents the results of the modified data structure of the individual texts, concentrating on letters only.

Table 2
Samples from Best (2005)

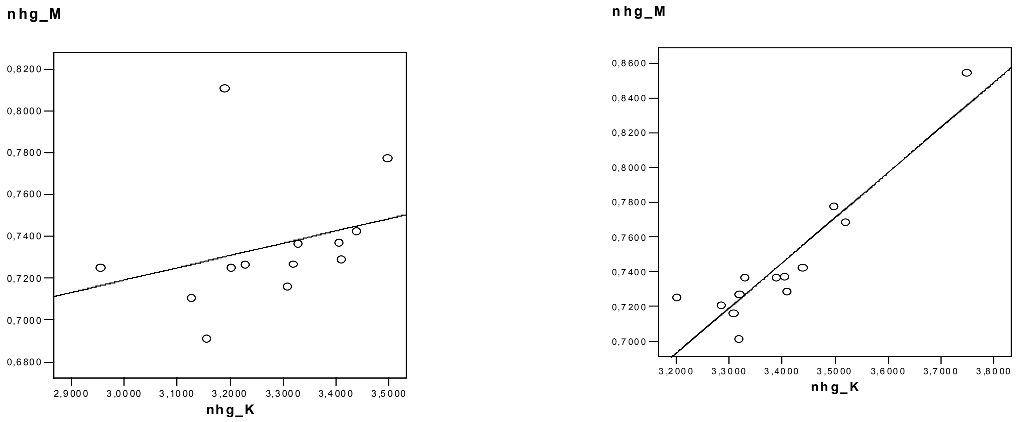
Nr.	Text	*N*	*I*	Basis
1	H. Pestalozzi: Hühner, Adler und Mäuse	675	27	Q,X,Y do not occur
2	G.A. Bürger: Münchhausen	137476	30	
3	G.A. Bürger: Lenore	6215	27	Q,X,Y do not occur
4	G. Büchner: Lenz	42608	30	
5	G. Büchner: Hessischer Landbote	21452	30	
6	K. May: Winnetou I	777361	32	É, Ñ; Ae -> Ä, Oe -> Ö, Ue -> Ü
7	F. Kafka: Die Verwandlung	99559	30	X = 0
8	F. Kafka: Der Prozeß	361848	30	
9	G. Vesper: Fugen	6259	27	Q,X,Y do not occur
10	O. Jägersberg: Dazugehören	40977	30	Y = 0
11	J. Joffe: Nach dem Bruderkrieg	6091	30	
12	R. Hoppe: Das gierige Gehirn	20075	30	
13	Schönpflug (1969)	99984	29	Without: ß
14	K.-H. Best: Wissenschaftliche Prosa	179922	30	

Achieving a systematic approach would consequently require a unitary treatment of the data to be analyzed. In principle, there are two options which shall both be pursued in our re-analysis:

1. the first alternative restricts the data sets to the analysis of those letters which occur in the relevant material, thus simply ignoring "missing" letters;
2. the second alternative assumes a given system to have a fixed inventory size and consequently integrates empty classes with frequency $f_i = 0$ into the data sets.

Whereas the first approach, which tolerates varying inventory sizes, thus meets the desires of a given "text", the second procedure is oriented to a system's needs. It will be interesting to compare the parameter behavior under these two conditions. As a matter of fact, this comparison must concentrate on the relation between parameters *K* and *M*, since the study of *K* and *I* makes no sense with fixed inventory size.

Figure 4 shows the differences for both conditions; although, after all, only 5 of the 13 samples have an altered inventory size, the differences are extremely clear. Figure 4a shows the effect of taking inventory size into consideration not on the basis of the given system, but on the observed realizations in each individual text: Under this condition, the linear trend to be observed in Figure 4a (with *I* = 30) is clearly disturbed; obviously the variation of *I* (directly reflected in parameter *n* of the negative hypergeometric distribution) also affects the parameter values *K* and *M* and thus disturbs, or even prevents, their behavior from being systematic and, as a consequence, amenable to a reasonable interpretation.



(a) *K* and *M* (I = varying) $r = 0.29$ ($p = 0.33$)

(b) *K* and *M* ($I = 30$) $r = 0.92$ ($p < 0.001$)

Figure 4. Correlation between parameters *K* and *M* (only letters)

In contrast to this, fixing the inventory size at $I = 30$, yields a clear correlation between parameters *K* and *M* ($r = 0.92$, $p < 0.001$) – cf. Figure 4b. The parameter values thus obtained are shown in Table 3, asterisks indicating diverging samples.

Table 3
Fitting results for two conditions

	$I = 30$				I = varying		
	K	*M*	*C*		*K*	*M*	*C*
1	3,3071	0,7163	0,0110	1	3,3071	0,7163	0,0110
*2	3,7480	0,8546	0,0125	2	3,1886	0,8109	0,0074
3	3,4083	0,7289	0,0100	3	3,4083	0,7289	0,0100
*4	3,3167	0,7016	0,0073	4	3,1549	0,6912	0,0058
5	3,4381	0,7427	0,0102	5	3,4381	0,7427	0,0102
*6	3,2839	0,7210	0,0100	6	3,1257	0,7108	0,0084
7	3,3184	0,7269	0,0104	7	3,3184	0,7269	0,0104
*8	3,5189	0,7688	0,0092	8	2,9541	0,7251	0,0074
9	3,4964	0,7775	0,0066	9	3,4964	0,7775	0,0066
10	3,4038	0,7372	0,0058	10	3,4038	0,7372	0,0058
11	3,3278	0,7366	0,0052	11	3,3278	0,7366	0,0052
*12	3,3886	0,7366	0,0062	12	3,2268	0,7265	0,0046
13	3,2002	0,7254	0,0079	13	3,2002	0,7254	0,0079

This finding for the first time documents systematic parameter behavior not only for Slavic, but also for German letter frequencies.

The next step is to investigate whether this systematicity is eventually bought at the cost of worse fitting results. To be sure, a better fitting result alone should not be the decisive factor in favoring one of the two options – in any case, a procedure which can be theoretically motivated is preferable.

Yet, as the analysis shows, the fitting results are almost equally good under both conditions; on average, the discrepancy coefficient is $C = 0.009$ for the “fixed condition”, and thus, only slightly worse than the “system condition” with $C = 0.008$. Comparing the fitting results for both conditions with the non-parametric Mann-Whitney-U-Test, differences between both conditions turn out to be not significant ($z = -0.85$, $p = 0.42$). This result is reflected by Figure 5, which contains an error bar diagram for the *C* values.

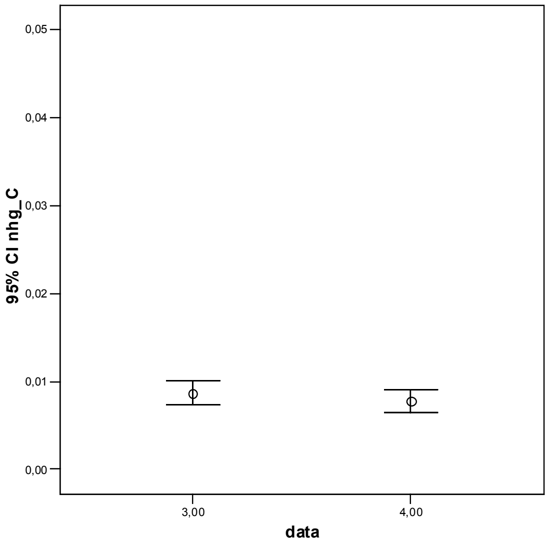


Figure 5. Error Bar Diagrams for C values

5. Summary and Perspectives

The results obtained in this study are a clear indication that the frequency of letters are regularly organized. Given the finding that the negative hypergeometric distribution has been shown to be an adequate model for both several Slavic languages and German, the present study provides additional evidence that the parameter behavior follows clear rules as well. As has been shown, however, this is only the case if the analysis is based on the system’s inventory size rather than on the number of classes observed in the individual samples.

Additional interpretations of the concrete parameter values must be left for future research. As has been argued elsewhere (Grzybek 2007), it seems that, in addition to inventory size, it is the mean of the distribution on the one hand, and the relative frequency of the most frequent class on the other, which rule the system’s overall behavior. It seems likely that estimating the parameter values of these statistical characteristics results in easy point estimations, which would explain the frequency behavior of letters; however, a definitive answer to this question must be left to the results of ongoing research.

References

Altmann, Gabriel; Köhler, Reinhard (1996). „Language Forces" and synergetic modeling of language phenomena. *Glottometrika 15, 62-76.*

Best, Karl-Heinz (2005). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11*, 9-31.

Grzybek, Peter (2005). A study on Russian graphemes. In: Toporov, V. N. (ed.), *Jazyk – ličnost’ – tekst. Sbornik statej k 70-letiju T.M. Nikolaevoj*: *237-263*. Moskva: Jazyki slavjanskich kul’tur.

Grzybek, Peter (2007). What a Difference an ‚E' Makes: Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen. In Deutschmann, P. (ed.), *Kritik und Phrase*. Wien. [In print]

Grzybek, Peter; Kelih, Emmerich (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 31*, *131-162*.

Grzybek, Peter; Kelih, Emmerich; Altmann, Gabriel (2004). Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung. In: *Anzeiger für Slavische Philologie 32*, *25-54*.

Grzybek, Peter; Kelih, Emmerich (2005a). Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens. *Glottometrics 9*, *62-73*.

Grzybek, Peter; Kelih, Emmerich (2005b). Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph. In: Altmann, G., Levickij, V., Perebejnis, V. (eds.), *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics*: *159-179*. Černovici: Ruta.

Grzybek, Peter; Kelih, Emmerich (2005c). 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, Peter; Kelih, Emmerich; Altmann, Gabriel (2005a). 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*.

Grzybek, Peter; Kelih, Emmerich; Altmann, Gabriel (2005b). Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen In: Kozmová, R. (ed.), *Sprache und Sprachen im mitteleuropäischen Raum. Vorträge der internationalen Tagung der internationalen Linguistik-Tage. Trnava 2005*: *641-664*. Trnava: GeSuS.

Grzybek, Peter; Kelih, Emmerich; Altmann, Gabriel (2007). Graphemhäufigkeiten im Slowakischen.. In: Nemcová, E. (ed.), *Philologia actualis slovaca*: Trnava. [In print]

Grzybek, Peter; Kelih; Emmerich; Stadlober, Ernst (2006b): Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik. *Anzeiger für Slavische Philologie 34, 41-74*.

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

Wimmer, Gejza; Altmann, Gabriel (2005). Unified derivation of some linguistic laws. In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807.* Berlin / New York: de Gruyter.

Wimmer, Gejza; Altmann, Gabriel (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, Peter (ed.): *Contributions to the Science of Text and Language. Word Length Studies and Related Issues: 329-337*. Dordrecht, NL.

Glottometrics 15

2007
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**

Herausgeber – Editors

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. –15 (2007-04-136) –. – Lüdenscheid: RAM-Verl., 2007
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 15 (2007)
ISSN 1617-8351

Contents