Studies
in Quantitative Linguistics
5

Issues

in

Quantitative Linguistics

edited by

Reinhard Köhler

RAM - Verlag

Issues in Quantitative Linguistics

edited by

Reinhard Köhler

2009
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Ján Mačutek (jmacutek@yahoo.com)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Preface

The present volume is a collection of recent papers on diverse linguistic and textological topics but with a common epistemological and methodological background. They contribute to the field of quantitative linguistics either by results of the application of quantitative approaches to interesting problems or by presenting new ideas and methods. A number of these contributions are versions of papers presented on occasion of the 5th Symposium on Quantitative Linguistics in Trier, Germany, December 2007.

Two of the papers are devoted to research in the field of stylistics. *Sergey Andreev* presents an investigation of an author's (Lermontov's) style with the emphasis on its development over the years during his short life (1814-1841). 35 texts including 25 poems were selected as empirical material; characteristics from several levels of linguistic analysis (morphology, syntax, rhythm, rhyme) serve as style indicators. Andreev arrives at the conclusion that two main periods in Lermontov's life can be determined, and central phases can be differentiated from peripheral ones when the individual texts are attributed to the periods. On data from five Modern Greek novels written by four authors, *George Mikros* conducts an authorship attribution experiment comparing different sets of stylometric characteristics. Besides common indicators, Mikros uses the most frequent functions words and the most distinctive author-specific words. His results yield convincing superiority assessments.

The application of Multidimensional Scaling (MDS) to geolinguistic data, as presented and illustrated by *Sheila Embleton, Dorin Uritescu, and Eric Wheeler* allows, when integrated into a software package and a corresponding data-base to select, search, count, view, edit, and analyse the data according to the researcher's interest. MDS, one of the statistical methods to reduce the number of dimensions of multidimensional data (in this case to just two dimensions), was implemented by the authors in their Romanian Online Dialect Atlas. Their presentation of their MDS function, which can be used for conveying an overview of the linguistic distances among locations with related dialects, gives the reader an impression of the explorative power of the approach. Another paper on a geolinguistic topic is the one presented by *Thomas Zastrow* and *Erhard Hinrichs*. They compare two approaches to computational dialectometry, which they characterize as an information theoretic approach and a vector-based one, on a Bulgarian data set. They, too, illustrate their work and show that both methods yield the same results, thus corroborating the approaches in an impressive way.

Slavic letter frequencies form the topic of *Peter Grzybek's, Emmerich Kelih's, and Ernst Stadlober's* research, which systematically corroborates the hypothesis that these frequencies are distributed according to the negative hypergeometric distribution (NHG). A surprising result of the comparative studies on data from five Slavic languages is the dependency of the NHG parameters on

language-specific factors as well as on interlingual ones. The authors are able to single out individual factors and to show their influence on parameter behavior.

Quantitative studies in linguistics are almost exclusively based on a ”bag-of-words” model, i.e. they disregard the syntagmatic dimension, the arrangement of units in higher units or on higher levels and in the course of the given text. The paper contributed by *Reinhard Köhler and Sven Naumann* shows how motifs, the recently introduced sequences of linguistic features, can be used for the analysis of texts also on the basis of clause properties. A second aim of this paper is the development of an algorithm for the automatic identification and segmentation of clauses in German sentences as a prerequisite for the study of linguistic mass data on this level. Another study on linguistics motifs is contributed by *Ján Mačutek*. He devotes his paper to the aspect of motif richness in analogy to vocabulary richness, a very popular problem in some branches of QL, based on word lengths motifs with length measured in terms of the number of syllables. The data have been taken from two Slovak texts.

An experiment is reported by *Adam Pawłowski, Maciej Piasecki, and Bartosz Broda*. They compared Michael Fleischer’s word profiles – collective symbols distilled from surveys – to profiles generated by automatic extraction from a corpus. The project explores in how much the results of a distributional extraction from text data match with semantic information given by human subjects as obtained in surveys and word priming experiments.

Two papers are devoted to research in the area of morphology. *Olga Pustylnikov and Karin Schneider-Wiejowski* address the phenomenon of productivity in derivational morphology from the point of view of its quantification. They evaluate three quantitative approaches proposed in the literature to measure productivity of German noun suffixes. In addition, they apply a decomposition algorithm used in a multi-agent simulation to identify productive suffixes. As opposed to most other studies on morphological productivity, the authors enclose in their empirical material written texts as well as oral speech. *Petra Steiner* scrutinizes an aspect of inflectional morphology. She deduced, in analogy to models of semantic diversification known from G. Altmann’s works, hypotheses for the distribution of the complexity of inflectional paradigms and tests them with four different measures on data from the Icelandic language. *Relja Vulanović* investigates another aspect of grammar, viz. properties of parts-of-speech systems. Flexible parts-of-speech systems are analyzed from the point of view of grammar efficiency. Seventeen linguistic structures are considered, most of them corresponding to natural languages described in typological samples. Vulanović shows that grammar efficiency of natural languages is well below the theoretically possible maximum.

The diachronic perspective is reflected in *Shoichi Yokoyama’s* and *Haruko Sanada’s* paper on language change. They introduce the models of language change known from QL research (Altmann’s Piotrowski Law) and illustrate them on hypothetical data. Their specific point of view as presented in the paper is a psychological view on the mechanisms behind the process, i.e. they assume an

intra-personal variable as a critical factor which determines the dynamics of the phenomenon.

Jan Králík's "contemplation" discusses the concept of infinity from different points of view. This discussion forms the background of his methodological and epistemological argumentation around the question as to if, when and in how far text and corpus studies can be compared to each other. Arguments from the theory of probability as well as theoretical and empirical findings in quantitative linguistics are taken into account.

I would like to thank the contributors for their co-operation; special thanks are due to Gabriel Altmann for his invaluable support and critical reviews.

Trier, December 2009 RK

Contents

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 17-33.*

Slavic Letter Frequencies:
A Common Discrete Model and Regular Parameter Behavior?

Peter Grzybek
Emmerich Kelih
Ernst Stadlober

Letter Frequencies and Frequency Models in the Context of Dynamic and Synergetic Linguistics

In the framework of quantitative approaches to language, so-called "low-level" units of language – e.g. letters, phones, phonemes, etc. – have always played a major role, from the early beginnings of this discipline on. Whereas earlier attempts in this field, which were mainly mere letter or sound statistics and the like, were related not only to linguistic problems, but also to concrete practical or technical issues of different kinds (cf. Grzybek 2006, Grzybek & Kelih 2003), recent studies are much more theory-based and, in fact, theory-oriented. A major reason for this development can be seen in the rise of synergetic linguistics (cf. Köhler 2005); in this context, letters (and other "low-level" units) can be seen as linguistic entities which form, or rather are part of systems, the characteristics and needs of which seem to be quite easy to survey as compared to more complex systems, where one is concerned with multi-faceted needs and multi-level influences. Therefore, it seems likely and reasonable, that any understanding of these allegedly less complex systems will yield deep insight into the dynamic mechanism of linguistic systems, in general; seen from this perspective, the study of letter frequencies clearly goes beyond simple analyses on something like a linguistic playground, and it represents much more than a methodological test case, but is an important and valuable scholarly object in its own right, contributing to a deeper understanding of the dynamics of linguistic systems.

Since general characteristics of letter inventories and frequency organization are of primary relevance here, the specific frequency of individual letters fades into the background. Instead, the question in how far the system-bound organization of a letter frequency distribution underlies general regularities comes to the fore. To this end, the frequency distribution is transformed into a (descending) order, where the frequency of the most frequent letter is assigned to the first rank, and the most infrequent letter to the last rank. The crucial question then concentrates on the point whether the frequencies exhibit a particular relation, or proportion and how these relations can best be described by a theoretical model.

The theoretical background of this procedure has repeatedly been described in recent years (Grzybek & Kelih, 2003; Grzybek, Kelih & Altmann 2004); a redundant description of the method can be abandoned here. The usual assumption in this context is that the probability of a given class with value *x* or rank *r* is proportional to the next lower class (i.e., *x–1* or *r–1*, respectively). Based on this general assumption (cf. Altmann , Köhler 1996) one may establish the difference equation

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

the concrete solution of which depends on the function *g*(*x*). In the past, even relatively simple functions *g*(*x*), usually rational functions, have yielded concincing results for the frequency analysis of linguistic units from different levels. More recently, Wimmer & Altmann (2005, 2006) have generalized this approach, and within this generalization, many distributions relevant for linguistic modeling may be sub-summarized under a common "linguistic roof". Without going into details here, let it suffice to say that this approach has also been successfully applied in systematic analyses of letter frequencies for various Slavic languages.[1]

One major objective of all these studies has been to systematically test previously discussed frequency distribution models with consistent material across different languages. Taking into account different "philosophies" of writing systems, our intention is not to find an overall valid, "universal" model for letter frequencies. Rather, the concentration on different Slavic languages offers the chance to study typologically similar languages from one and the same linguistic family, i.e. languages which share some general common straits, but still display some variation; this might shed light on some factors influencing the system-related behavior of this linguistic level, and one should expect that, given an adequate model common to these languages, relevant changes might result in some interpretable parameter behavior yielding deep insight into the synergetic organization of this level.[2]

Thus far, only selected languages have been analyzed, and the results obtained should be taken with a pinch of salt. Anyway, as a first result, it turned out that most of the models discussed in the past turned out to be inadequate; only

[1] For Russian see Grzybek & Kelih (2003), Grzybek, Kelih & Altmann (2004), Grzybek, Kelih & Altmann (2005a) and Kelih (2007); for Slovak see Grzybek, Kelih & Altmann (2005b) and Grzybek, Kelih, Altmann (2006); for Ukrainian see Grzybek & Kelih (2005a), and for Slovene see Grzybek, Kelih, Stadlober (2006).

[2] In detail, these models are the zeta distribution, the Zipf-Mandelbrot distribution, the geometric distribution, the Good distribution, the Whitworth distribution, the negative hypergeometric distribution.

one model, the negative hypergeometric (NHG) distribution, could be shown to suitable for letter frequencies of the languages studied thus far.[3]

As compared to all other distribution models, the NHG distribution – which shall be presented in detail below – has the most parameters; it goes without saying that the more parameter a distribution model has, the more flexible it is. The parameters of a given distribution have to be estimated such that the model yields the best fit to the data under study. In former times, this estimation has been done by different estimation methods, where estimated values were determined with regard to theoretical characteristics of the given model. Today, this process of parameter estimation is increasingly, if not exclusively, done by special software tools: parameters are optimized via iterative procedures to obtain minimal deviations between theoretical and observed values.[4] As a matter of fact, parameter estimation is first and foremost a method to find the optimal parameter values. Then the corresponding model values have to be tested statistically for goodness of fit. Yet, fitting of the distribution model is of course not the ultimate aim; rather, this is one important step in the course of a quantitative linguistic study, which should, at the end, lead to some qualitative interpretation of the results obtained. At this phase the crucial transition from the *discovery* and *description* of particular regularities to their *interpretation* and eventual *explanation* should take place. This clearly defined step is not self-evident in qualitative linguistics, what sufficiently characterizes the latter's scientific status… One important step in this transition would be, of course, the availability of some interpretation of the parameter behavior. However, also in quantitative linguistics, ultimately striving at theoretical explanations, parameter interpretations have hardly ever been achieved and remain one of the crucial objectives of research.

This intention is the starting point of the present study: Based on the observation that obviously, for the description of Slavic letter frequencies, a complex model such as the NHG distribution with its three parameters *K*, *M*, and *n* to be estimated (for details see below) is needed, an attempt shall be made to approach at least some partial explanation of parameter behavior across the languages studied. This endeavor might then be considered to be successful if the

[3] Interestingly enough, the NHG distribution has been proven to be adequate not only for Slavic languages, but for German, as well, cf. Best (2004/05, Best 2005, Grzybek 2007a,b); further details must remain unconsidered, here.

[4] For reasons discussed elsewhere in detail, we do not work with continuous models and curves, here (as to this line of research, cf. the recent work by Kelih 2009), but with discrete frequency models, only. In the studies reported here all relevant approaches thus far pursued in studies on letter frequencies, have been tested for their adequacy. The goodness of fit has been tested with statistical procedures; first and foremost, this has been done by the chi-square test. Since the latter increases linearly with increasing sample size (resulting in significant deviations even in case of good fits), it is more reasonable to use the discrepancy coefficient $C = \chi^2/N$. Values of $C < 0.02$ are then interpreted as an indication of a good fit, values of $C < 0.01$ of a very good fit.

systematic of rank frequency behavior of Slavic letters might be grasped not only within each of the individual languages, but also in comparison across languages. In case this attempt should turn out to be successful, this would be an important step in explaining the necessity of such a complex model.

With these perspectives in mind, it seems reasonable to shortly summarize the state of the art and to present the languages and material analyzed, before delving into further details.

0. Previous Studies on Slavic Languages

Systematic studies on grapheme frequencies have been reported with regard to four Slavic languages: Russian, Ukrainian, Slovak, and Slovene, thus covering inventory sizes (I) in the interval of $25 \leq I \leq 46$, the minimum of 25 representing Slovene, the maximum of 46 representing Slovak (including diagraphs):

1. **The Slovenian data** are taken from Grzybek, Kelih & Stadlober (2006). The experimental framework of this paper may be summarized as follows: 30 individual texts from different text sorts (masters theses, journalistic comments, sermons, private letters, literary prose and scholarly articles) were analyzed. Across all 30 samples, the discrepancy coefficient for the NHG-Distribution was in the interval $0.022 \geq C \geq 0.0055$; for 26 of the 30 individual analyses, we had $C < 0.02$; for 6 of them even $C < 0.01$. Thus the NHG distribution seems to be an adequate model for the Slovenian grapheme frequencies analysed.
2. **Russian grapheme frequencies** were examined in Grzybek, Kelih, Altmann (2005a), involving 30 complete texts. Again in six different text sorts (drama, stories, poems, private letters, novel chapters and novel in verse) the grapheme frequencies were studied under two different conditions: (a) with the letter ‚ë' as a letter in its own right and without it (represented as 'e' instead) – inventory size thus changing from $I = 32$ to $I = 33$. This specific design did not primarily intend to make a „political" statement as to the status of letter; rather, it was meant to be a detailed analysis of the relevance of inventory size for grapheme studies. As a result it turned out that, apart from a systematic displacement (and in fact no significant differences) of entropy and repeat rate values, again the NHG distribution was a good model under both conditions (with $C < 0.02$ in 59 of 60 samples). With condition $I = 32$ – used for our re-analysis of the parameters below – 21 texts showed $C > 0.001$ and for the remaining rest $C < 0.002$ was obtained. In general, once again the NHG distribution, turned out to be a suitable model for grapheme frequencies in Russian.
3. **Slovak** grapheme frequencies were studied on the basis of 30 texts (literary prose, diploma theses, journalistic comments, fairy tales and "technical" texts) where again the NHG distribution turned out to be the

only adequate model (Grzybek, Kelih & Altmann 2005b und 2006); for Slovak, too, this holds true for two conditions, differing with regard to inventories: taking the three digraphs „dz“, „dž“ and „ch“ as separate units in their own right, inventory size is $I = 46$, otherwise $I = 43$. With $I = 46$, 25 of the 30 analyses yielded a fit of $C < 0.02$. The fitting results under condition $I = 43$ are as follows: 28 of 30 texts had $C > 0.02$; 10 texts had even $C > 0.01$. Two outliers ($C > 0.02$) can be explained due to the relatively small sample sizes of 562 and respectively 445 graphemes. Nevertheless, the NHG distribution fits well for Slovak grapheme frequencies too.

4. Finally, **grapheme frequencies of Ukrainian** were studied by Grzybek & Kelih (2005a). The study included 30 texts (drama, journalistic texts, poems, literary prose and scientific texts); inventory size here amounts to $I = 33$ (not counting the inverted comma as a separate grapheme). Again, the NHG distribution was shown to be an adequate model, with $C < 0.02$ in all 30 texts and even $C < 0.01$ for twenty of them.

1.1. Results: Details

Summarizing, one can say that the grapheme ranking behavior can be grasped by one type of model across the four languages studied: This model is the NHG distribution, in its 1-displaced form (since ranking usually starts with rank 1):

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

Taking n as one of three parameters of the NHG distribution fixed at $n = I-1$ (since the support of the NHG distribution is limited by $n+1$), only K and M remain as two free parameters to be estimated. With regard to the individual analyses, both parameters may differ for two reasons: first, they may differ within a given language (due to a “natural” variance of frequencies), and second, they may differ across languages (obviously due to some specific ranking behavior). To systematically analyze the parameter behavior of K and M, and to find possible general tendencies, it seems reasonable to calculate mean values of K and M within each of the given languages, along with 95% confidence intervals. Table 1 represents the corresponding values: in addition to the number of samples analyzed (n), information is given as to inventory size (I), as well as to mean values, upper and lower limits of the confidence intervals for K and M.

Table 1
Parameter Behavior of *K* and *M* in the Languages Analyzed

	n	*I*	$\overline{K}$	K_u	K_o	$\overline{M}$	M_u	M_o
Slovene	30	25	2.96	2.91	3.00	0.8351	0.8263	0.8439
Russian	30	32	3.14	3.10	3.18	0.8096	0.7990	0.8202
Ukrainian	30	33	2.96	2.92	3.01	0.8203	0.8082	0.8324
Slovak	30	43	4.07	4.00	4.14	0.8546	0.8389	0.8703

Based on these findings, first attempts have been undertaken to check the behavior of parameters *K* and *M* for regularities and look for possible interpretations (cf. Grzybek, Kelih 2005b; Grzybek, Kelih, Altmann 2005a). In these attempts, it has first been argued on a direct dependence of parameter *K* on inventory size *I*; as a consequence, the interpretation of *K* would be possible across languages. As compared to this, it has been argued in favor of a relation between parameters *K* and *M* in form of a linear relationship, though not across languages, but within each of the given languages. In other contexts (Grzybek 2007a,b) additional interpretations have been considered e.g. the possibility that parameter *M* may be related either to the first frequency (P_1) of a given distribution, or to its mean rank (m_1).

For the time being, these far-reaching perspectives will not be further pursued, here; instead, the observed dependence of *M* and *K* is analyzed in more detail, and a statistical test is presented which may be useful in the given situation.

Figure 1 demonstrates the relation between the two parameters for the four samples described above.

There is only a weak dependence of *M* on *K* across languages, but a significant linear relationship ($p < 0.001$) within each of the languages, *M* increasing with an increase of *K*, and with correlation coefficients *r* ranging from 0.79 to 0.89 and in all cases.

A comparison of the parameter behavior between the different languages shows that the overall tendency seems to be almost identical, displaying approximately parallel slopes of the four regression lines.

A closer inspection of Figure 1, however, displays two remarkable deviations from expectance:

1. for Ukrainian, parameter *K* seems to be smaller than expected (i.e., not in line with the inventory size interpretation);
2. the regression line for Slovene deviates from the scheme, despite its overall accordance with the general parallel tendency, being characterized by an intersection with the regression line of the Ukrainian data.

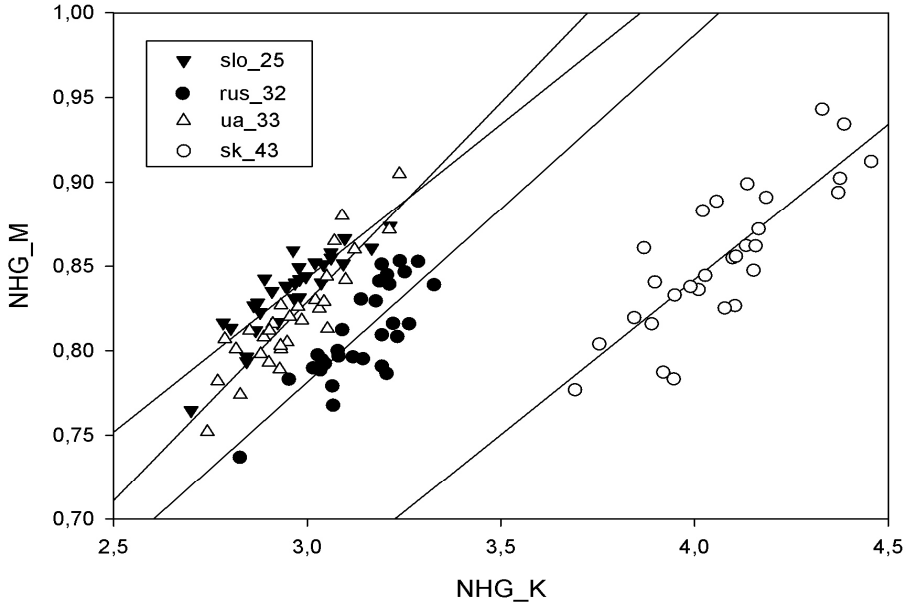


Figure 1. Dependence of parameter *M* on *K* for Slovenian, Russian, Ukrainian, and Slovak grapheme data

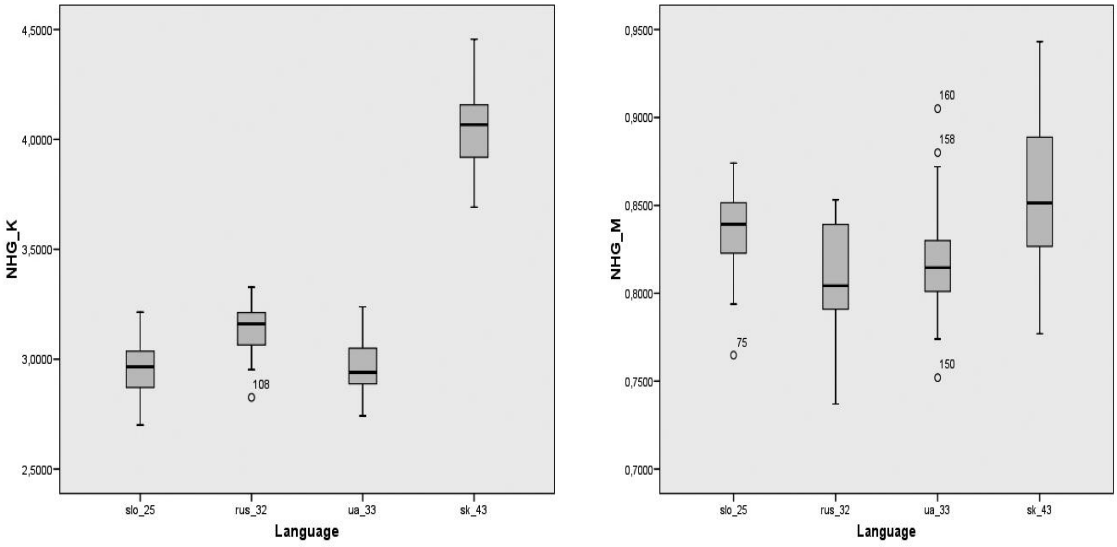
Focusing on the relation between *M* and *K*, only, we neglect the first problem in the given context and concentrate on the second issue. Thus, first stating the overall aptness of the NHG distribution for modeling Slavic letter frequencies, and second observing a general tendency in the behavior of parameter *M*, we can turn to the more specific question as to the observed deviations from the established rule.

1.2. Outliers and Extreme Values

A common first step in explaining the observed deviations from general parameter behavior can be seen involves an analysis of possible outliers and extreme values, which are eliminated from the analysis. This is usually be done be reference to the so-called interquartile range (IQR), which comprises the central 50% of all observations. Outliers and extreme values are defined as cases, for which the difference to the upper and lower limit of the IQR is more than 1.5 times (or 3 times, respectively) as large as the IQR.

The analysis can be illustrated by box plots, in which outliers can easily be detected and identified: they are located beyond or below the upper or lower line, which is defined by a concrete value of the given data set, and is maximally 1.5

times as large as the IQR – if there are no outliers, they are represented by the maximal and minimal values of the given sample. Figures 2a and 2b represent the four box plots for the parameter values of *K* and *M*.



(a) Parameter *K* **(b)** Parameter *M*

Figure 2. Box plot series

Indeed there are some outliers which can easily be identified. In case of parameter *K*, this is only one of the Russian private letters (# 260). In case of parameter *M*, we are concerned with four outliers, one from the Slovene data (# 22), and three from the Ukrainian data (# 332, 340, 342). Eliminating these outliers from the analysis and submitting the data again to a study of parameter behavior results in an only slightly changed picture of the regression lines, as illustrated by Figure 3.

As can be seen, the regression line for the Slovene data is still characterized by an intersection with the regression line of the Ukrainian data, but now this intersection is far away from all observed data points. Table 2 represents the regression equation for all four languages (after elimination of the outliers); these regression lines follow the equation $y = a + bx$ (in our case we have $M = a + bK$). Inventory size is denoted by *I*, sample size by *n* (after elimination of outliers);

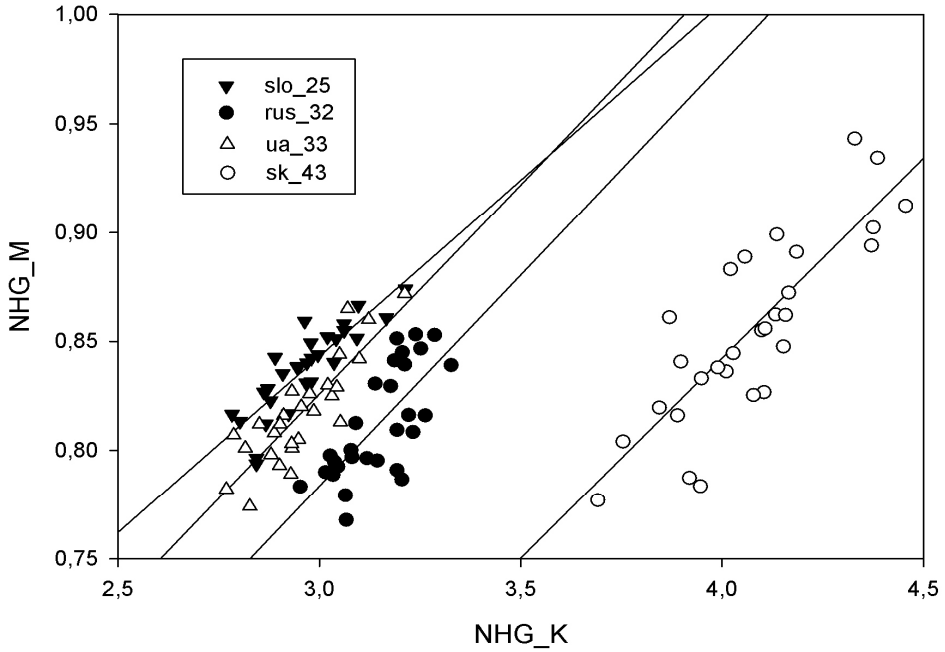


Figure 3. Relation between parameters *K* and *M* (after elimination of outliers)

Table 2
Regression coefficients: $M_i = a_i + b_i \cdot K_i$ and correlation coefficients *r*

	I	*n*	*b*	*a*	*r*
Slovene	25	29	0.1620	0.3571	0.86
Russian	32	29	0.1941	0.2013	0.72
Ukrainian	33	27	0.1921	0.2494	0.85
Slovak	43	30	0.1840	0.1067	0.83

Since it is the regression line for Ukrainian, which intersects with the Slovenian one, it is reasonable to test the difference between the two regression coefficients (slopes b_1 and b_2) for significance. For linear relations, this can be done by reference to a *t*-test statistic

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

with $DF = n_1 + n_2 - 4$ degrees of freedom and $Q_x = \sum (x - \bar{x})^2$.

As a result, the comparison of the two regression coefficients b_1 for Slovene and b_2 Ukrainian shows the difference to be not significant, with a value of $t = 1.01$ and $DF = 52$ degrees of freedom ($p = 0.32$). This result naturally leads to a simultanous comparison of all four regression lines, rather than only two of them. Given there is no significant deviation from parallelism, this would yield a regression model common to all four samples studied.

1.3. A common regression model for Slavic letter frequencies?

With regard to a possible uniformity of the tendencies and, as a consequence, a common regression model, one may ask the specific question if the dependence of parameter *M* on *K* shows an identical trend for the four languages under study. This leads to the question of testing the differences between the regression coefficients and the parallelism of the regression lines for significance. An adequate procedure to test this is the multiple partial *F*-test; usually, this test is applied with regard to multiple linear regressions (cf. Kleinbaum et al. 1998) when the question of possible additional contributions of independent variables, which are not (yet) included in a given model, is at stake. The *F*-test thus tests the effect of expansion of a given model by the simultaneous addition of two or more variables. In its complete form, such a multiple model has the following form:

$$Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \beta_1^* X_1^* + ... + \beta_k^* X_k^* + \varepsilon \quad (4)$$

Here, *Y* is the dependent variable, α is the regression constant, and ε is a random error; *X*i and X_i^* are the independent variables, β_i and β_i^* the regression coefficients. The null hypothesis (H_0) to be tested includes the assumption that $X_1^*, X_2^*, ..., X_k^*$ do not significantly contribute to the prediction of Y, when X_1, $X_{2...}$ X_k are already included in the model; thus, for the complete model we have $H_0 : \beta_1^* = \beta_2^* = ... = \beta_k^* = 0$.

From this (second) formulation the reduced model under H_0 is:

$$Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon \quad (5)$$

Thus, the variance (SSq_{reg}) explained by the model becomes larger by the addition of X_i^*; now, the following F statistics can be calculated:

$$F = \frac{\left[SSq_{reg}(complete\ model) - SSq_{reg}(reduced\ model) \right] / k}{SSq_{res}(complete\ model) / \left(n - p - k - 1 \right)} \quad (6)$$

In (6), SSq_{reg} denotes the variance (i.e. the sum of the squared deviations: sum of squared effects) of the complete or the reduced regression models, and SSq_{res} denotes the squared sum of the residuals of the complete model (sum of squared errors); n is the sample size, p is the number of regression coefficients in the reduced model, and k is the number of those regression coefficients which equal zero under the assumption of the null hypothesis (H_0).

In our case, the sample size is n = 115 texts (without outliers) from four sub-samples from four different languages (each with its own inventory size). Since the attribution to one of the languages is a nominal category, and since nominally scaled predictors cannot be directly introduced into a regression model, the relevant information has to be (re)-coded in a different manner. To this end, one introduces dummy coding: in this case, a variable is split into sub-variables (which are termed 'indicators') and coded dichotomously; each category is thus classified 'present' (1) or 'absent' (0). Membership of a given case within a given (sub)sample can thus be regarded as a dummy variable with the coding 0 and 1. The advantage of such a binary (0 vs. 1) coding is that dummy variables can be statistically treated like interval-scaled variables. A categorical variable with *k+1* values is thus transformed into *k* dummy variables each with two values 0 and 1. Since our variable "LANGUAGE" (with a given inventory size) has four categories, three dichotomous variables (D_1 to D_3) can be constructed which contain the same information as one categorical variable. In our case, we thus obtain the scheme represented in Table 3.

Table 3
Coding schema and dummy coding

	I		D_1	D_2	D_3
Slovene	*25*	*8*	0	0	0
Russian	*32*	*2*	1	0	0
Ukrainian	*33*	*5*	0	1	0
Slovak	*43*	*6*	0	0	1

Within this framework, our question as to the parallelism of regression lines turns out to be a special case of a more general situation: considering the regression lines to be parallel to each other if the predictive value of *Y* is not significantly changed by the addition of the additional variables, the described procedure as apt to be applied to this special case. In this case, the reduced model for *M* can be written as:

(7) $$M = \alpha + \beta_1 \cdot K + \beta_2 \cdot D_1 + \beta_3 \cdot D_2 + \beta_4 \cdot D_3 + \varepsilon \ .$$

This means that the regression lines of all four groups are parallel with identical slope β_1 (p = 4). The pre-conditions are thus fulfilled to make a comparison between the complete and the reduced model, with regard to variable 'LANGUAGE' in its re-coded (dummy-coded) form, by addition of the products of X_i and K as variables X_1^*, X_2^*, X_3^* to the dummy variables $X_1^* = KD_1$, $X_2^* = KD_2$, $X_3^* = KD_3$. Hence the full model can be written as

(8) $$M = \alpha + \beta_1 \cdot K + \beta_2 \cdot D_1 + \beta_3 \cdot D_2 + \beta_4 \cdot D_3 + \beta_1^* \cdot KD_1 + \beta_2^* \cdot KD_2 + \beta_3^* \cdot KD_3 + \varepsilon$$

In our case we are concerned with 7 parameters for the complete model (K, 3 dummy-coded variables, and 3 dummy products), for which we obtain the values SSq_{reg} = 0.0726 and SSq_{res} = 0.0229.

Interestingly enough, one obtains for the reduced model (7), which contains four variables, namely, the three dummy-coded variables in addition to K, a nearly identical value of SSq_{reg} = 0.0724. The error sum of squares (i.e., the sum of the squared deviations of the residuals) of the reduced model, too, is almost the same with SSq_{res} = 0.0231.

Thus, in this particular case, the assumption of parallelism seems likely to be confirmed by a statistical test, the F-test.

For the calculation of the F value we need the value of k (cf. (6)), which is obtained by the difference between the number of variables of the complete model (8) and that of the reduced model (5), in our case, $k = 7 - 4 = 3$, equivalent to the number of degrees of freedom for the numerator in (6). We also need the mean of the squared sum of residuals, which is represented by the quotient of the sum of the squared residuals (SSq_{res}) and the number of the degrees of freedom of the denominator, being calculated as $m = n - p - k - 1$; in our case, we thus have $m = 115 - 7 - 1 = 107$.

We now can calculate the F value as

$$F_{(FG_3=5, FG_2=107)} = \frac{(0.0726 - 0.0724)/3}{0.0229/107} = 0.31$$

With the given degrees of freedom, this F-value corresponds to a probability of p = 0.82, which is far from any statistical significance; as a consequence, we have to retain the null hypothesis ($H_0 : \beta_1^* = \beta_2^* = \beta_3^* = 0$), according to which the regression lines are parallel.

We can thus summarize that the dependence of parameter M on parameter K of the negative hypergeometric distribution behaves identically across the four Slavic languages studied which can be expressed as a common regression model. Within this model, the common regression coefficient (slope) is $\hat{\beta} = b \approx 0.1811$; accordingly, for the four languages under study parameter M can be estimated as

$\widehat{M} \approx \widehat{\alpha} + 0.18 \cdot K$; differences between the languages are a result of differences in the intercept $a = \widehat{\alpha}$ From this general model, the individual groups (i.e., the four languages each with their given inventory sizes) can be derived as special cases. Given the fact that the null hypothesis is to be retained, it is sufficient to do this with reference to the reduced model. Ignoring the error of estimation (ε), we obtain

Group 1:	$a + b_1 \cdot K$
Goup 2:	$(a + b_2) + b_1 \cdot K$
Group 3:	$(a + b_3) + b_1 \cdot K$
Group 4:	$(a + b_4) + b_1 \cdot K$

For our four languages we thus obtain the following special regression models

Slovene	25	$M = 0.18K + 0.3005$
Russian	32	$M = 0.18K + 0.2470$
Ukrainian	33	$M = 0.18K + 0.2826$
Slovak	43	$M = 0.18K + 0.1181$

Now, interpreting the intercepts as response variables of a regression model with inventory size *I* as the independent variable exhibits a clear tendency as illustrated in Figure 4. This tendency, based on four data points only, is not significant ($p = 0.07$; $r = 0.93$), however, obviously due to the deviating structure of the Ukrainian data; the analysis of more data from further Slavic languages will shed more light on this highly intriguing issue.

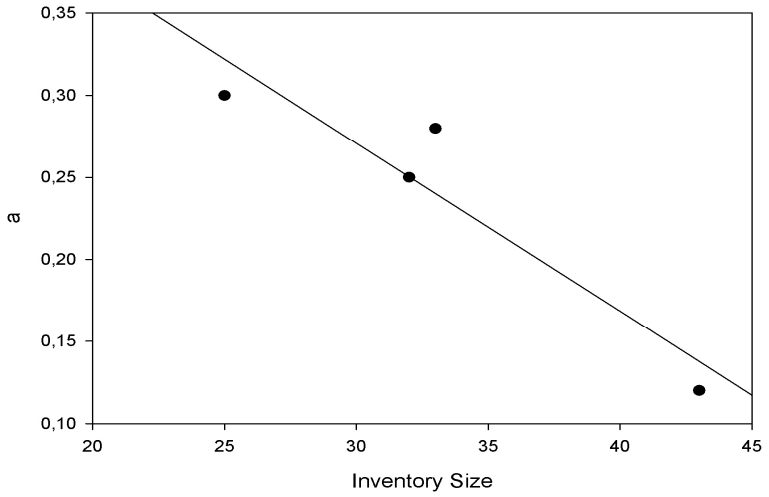


Figure 4. Relation between the intercepts of the regression model and inventory size *I*

With these findings, an important step seems to be made towards the intended analysis of the systematic parameter behavior, at least as far as parameter *M* of the NHG distribution is concerned: it turns out to be plausible that parameter *M* is closely related to parameter *K*, not across languages, but rather within a given language, only.

Yet, it is interesting to delve even deeper into the matter; a next logical step in this direction would be an answer to the question under what circumstances the outlined regression model is less effective than expected. To provide possible answers to this question, a next step might involve specific analyses of residuals, but this would clearly go beyond the scope of this paper.

2. Summary of Results and Future Prospectives

With regard to the findings reported above, we can summarize the most import results:

1. In systematic analyses of letter frequencies from four Slavic languages it could be shown that all can be adequately modeled by the NHG distribution, other models failed to be likely successful.
2. Parameter behavior of the NHG distribution seems to be highly regular; this regularity seems to be related to both language-specific and interlingual factors:
    a. the relevance of interlingual factors has already been discussed else where (cf., e.g., Grzybek & Kelih 2005); inventory size has been identified as a crucial factor influencing the distinction between languages; in this article, further arguments in favor of this notion have been brought forth, by showing that, at least for various Slavic languages, the relation between the parameters *M* and *K* of the NHG distribution follows a common linear regression model from which the individual languages may be derived as special cases;
    b. language-specific tendencies are expressed in parameter values, which lend themselves to discriminant analyses; also the specific relation between parameters *K* and *M* of the NHG distribution seems to be specific for individual languages.

    The study of additional languages, Slavic and non-Slavic, is necessary to gain more information on this specific situation. It seems plausible that, in addition to the above-mentioned language-specific and interlingual factors, also "local" factors may come into play, as can be seen in case of Ukrainian; here, additional analyses turn out to be necessary to grasp more exactly the boundary conditions of letter behavior. In this context, it has to be checked if the deviation of individual languages (as, e.g., Ukrainian in our case) may be caused by mere computational aspects of parameter estimation; to give an answer to this question, a qualitative parameter

interpretation is in order. We are still relatively far from this stage, but first analyses in this direction point at the importance of particular "first-order" characteristics such as inventory size (I), or first frequency (P_1), as well as of "second-order" characteristics, such as mean rank (m_1), entropy (H), repeat rate (RR), etc. – research in this direction is in progress now.

3. As this study shows, single corpus analyses of a given language cannot, as "representative" as they may be considered to be, uncover all mechanisms and processes at work in a language's dynamic system – in the given case we see that, within a language, there seems to be an intrinsic mechanism which synergetically regulates the dynamic balance of possibly contradictory forces, and which regulate the overall frequency behavior.
4. In order to identify trends, any sample must be checked for possible outliers and extreme values which eventually has to be eliminated from the analysis in order the trend to be uphold; in this respect, sample size, too, must be controlled to guarantee the statistical stability of tendencies (cf. Grzybek et al. 2009).
5. It is obvious that modifications of the model described above may be necessary when further languages (particularly from other than the Slavic family) will be taken into consideration; it may well turn out that the NHG model then soon turns out to be a special model relevant only for particular languages, or specific writing systems, etc.
6. In order to arrive at an explanation why the NHG is a suitable model for grapheme frequencies, its theoretical derivation must be carefully taken into account: since, in this case, it does not seem to make sense to interpret it in terms of an urn model, it might be reasonable, by way of an alternative, to derive the NHG rather as beta-binomial distribution, i.e., as a binomial distribution with its parameter p being variable and following a beta distribution: as Grzybek (in print) shows, this results in a new interpretation of the whole generating process. Seen from this perspective, a specific inventory size is not a "given" fixed starting point, but rather emerges as the diachronically motivated outcome of the dynamic speaker-hearer communicative interaction.

References

Altmann, G., Köhler, R. (1996). 'Language Forces' and synergetic modelling of language phenomena. In: Schmidt, P. (ed.), *Glottometrika 15, 62–76.* Trier

Best, K.-H. (2004/05). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 21–32*.

Best, K.-H. (2005). Buchstabenhäufigkeiten im Deutschen und Englischen. *Науковий вісник Чернівеского університета, віп. 231, 119–127.*

Grzybek, P. (2006). A very early Slavic letter statistic in the Czech journal *Krok* (1831): Jan Svatopluk Presl (1791-1849). *Glottometrics 13*, 88–91.

Grzybek, P. (2007a). What a difference an ‚E' makes. Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen. In: Deutschmann, P. unter Mitarbeit von P. Grzybek, L. Karničar, H. Pfandl (eds.), *Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag: 105–128.* Wien: Praesens.

Grzybek, P. (2007b). On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies. *Glottometrics 15,* 82–91.

Grzybek, P. (in print). Graphem- und Phonemstatistik: Inventare – Modelle – Zusammenhänge – Typologie. In: Kempgen, S., Berger, T., Gutschmidt, K., Kosta, P. (eds.), *Die Slavischen Sprachen. Ein internationales Handbuch zu ihrer Struktur, ihrer Geschichte und ihrer Erforschung. Bd. 2.*

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. (2005a). Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph." In: Altmann, G., Levickij, V., Perebejnis, V. (eds.), *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics: 159–179.* Černovci: Ruta.

Grzybek, P., Kelih, E. (2005b). 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., Altmann, G. (2004). Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie 32, 25–54*.

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

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

Grzybek, P., Kelih, E., Altmann, G. (2006). Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen. In: Kozmová, R. (ed.), *Sprache und Sprachen im mitteleuropäischen Raum: 661–684*. Trnava: GeSuS.

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

Grzybek, P., Mačutek, J., Stadlober, E., Wimmer, G. (2009). Sample size estimation in linguistics – A new approach. In prep.

Kelih, E. (2007). Häufigkeiten von Graphemen und Lauten: Zwei Ebenen – ein Modell? (Re-Analyse einer Untersuchung von A.M. Peškovskij). In: Grzybek, P., Köhler, R. (2006) (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann On the Occasion of His 75th Birthday: 267–277*. Mouton de Gruyter: Berlin – New York [= Quantitative Linguistics, 62].

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

Kleinbaum, D.G.; Kupper, L.L.; Muller, K.E. ([3]1998). *Applied regression analysis and other multivariable methods*. Pacific Grove: Duxbury Press. 3rd ed., rev.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch – An International Handbook*: *760–774*. Berlin u.a.: Walter de Gruyter [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Wimmer, G.; Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch – An International Handbook: 791–807*. Berlin u.a.: Walter de Gruyter..

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues: 329–335*. Dordrecht, NL: Springer.