# Towards a General Model of Grapheme Frequencies for Slavic Languages

Peter Grzybek[1] and Emmerich Kelih[2]

[1] Graz University, Austria, Dept. for Slavic Studies,
peter.grzybek@uni-graz.at,
[2] Graz University, Austria, Dept. for Slavic Studies,
emmerich.kelih@uni-graz.at,
WWW home page: http://www-gewi.uni-graz.at/quanta

**Abstract.** The present study discusses a possible theoretical model for grapheme frequencies of Slavic alphabets. Based on previous research on Slovene, Russian, and Slovak grapheme frequencies, the negative hypergeometric distribution is presented as a model, adequate for various Slavic languages. Additionally, arguments are provided in favor of the assumption that the parameters of this model can be interpreted with recourse to inventory size.

## 1 Graphemes and Their Frequencies

The study of grapheme frequencies has been a relevant research object for a long time. From a historical perspective, only a small part of the studies along this line have been confined to the mere documentation of grapheme frequencies, considering this to be the immediate object and ultimate result of research. Other approaches have considered the establishment of grapheme frequencies to be the basis for concrete applications. In fact, relevant studies in this direction have often been motivated or accompanied by an interest in rather practical issues such as, for example, the optimization of technical devices, the structure of codes and processes of information transfer, crytographical matters, etc.

A third line of work on grapheme frequencies has been less practically and more theoretically oriented. In this framework, research has recently received increasing attention from quantitative linguistics. As compared to the studies aoutlined above, the focus of this renewed interest has shifted: In a properly designed quantitative study, counting letters (or graphemes), presenting the corresponding absolute (or relative) frequencies in tables, or illustrating the results obtained in figures, is not more and not less but one particular step. In this framework, data sampling is part of the empirical testing of a previously established hypothesis, motivated by linguistic research and translated into statistical terms. The empirical testing thus provides the basis for a decision as to the initial hypothesis, and on the basis of their statistical interpretation one can strive for a linguistic interpretation of the results (cf. Altmann 1972, 1973).

Providing and presenting data thus is part of scientific research, and it is a necessary pre-condition for theoretical models to be developed or elaborated. As

far as such a theoretical perspective is concerned, then, there are, from a historical perspective (for a history of studies on grapheme frequencies in Russian, which may serve as an example, here, cf. Grzybek & Kelih 2003), two major directions in this field of research. Given the frequency of graphemes, based on a particular sample, one may predominantly be interested in

1. comparing the frequency of a particular grapheme with its frequency in another sample (or in other samples); the focus will thus be on the frequency analysis of individual graphemes;
2. comparing the frequencies of all graphemes in their mutual relationship, both for individual samples and across samples; the focus will thus be on the analysis and testing of an underlying frequency distribution model; this approach includes – if possible – the interpretation of the parameters of the model.

In our studies, we follow the second of these two courses. We are less interested in the frequency of individual graphemes. Rather, our general assumption is that the frequency with which graphemes in a given sample (text, or corpus, etc.) occur, is not accidental, but regulated by particular rules. More specifically, our hypothesis says that this rule, in case of graphemes, works relatively independent of the specific data quality (i.e., with individual texts as well as with text segments, cumulations, mixtures, and corpora). Translating this hypothesis into the language of statistics, we claim that the interrelation between the individual frequency classes is governed by a wider class of distributions characterized by the proportionality relation given in (1):

$$P_x \sim g(x)P_{x-1} \ , \tag{1}$$

relating a given class to previous classes, or by a partial sums relation, thus relating a class to the subsequent classes.

Thus, as opposed to studies focusing on the frequency of individual graphemes, the accent is on the systematic relation between the frequencies of all graphemes (or rather, the frequency classes) of a particular sample. Research thus is interested in the systematic aspects of frequencies, concentrating on the (relative) frequency of the most frequent grapheme, as compared to the second, third, etc. It is thus the study of the rank frequency distribution of graphemes in various texts and languages, which stands in the focus of attention. The objective is the theoretical modeling and mathematical formalization of the distances between the individual frequencies, irrespective of the specific grapheme(s) involved. Consequently, the procedure is as follows: If one transforms the raw data obtained into a (usually decreasing) rank order, and connects the data points with each other, one usually obtains not a linear decline, but a specific, monotonously decreasing (usually hyperbolic) curve. The objective then is to model the specific form of this curve, and to test, if the frequencies in different samples (i.e., the specific decline of the frequencies) display one and the same form, or not.

Thus far, convincing evidence has been accumulated to corroborate this hypothesis for three of the Slavic languages: Slovene, Russian, and Slovak. The

basic results have been presented in detail elsewhere – cf. Grzybek, Kelih, & Altmann (2004) for Russian, Grzybek & Kelih (2003) for Slovene, and Grzybek, Kelih & Altmann (2005a,b) for Slovak. The present contribution is a first attempt to arrive at some synopsis and to develop some generalizing conclusions. Therefore, it will be necessary to briefly present the results hitherto obtained by way of some summary, before we turn to a synopsis of these results, which will ultimately lead to some hypothesis for further studies.

## 2 A Model for Grapheme Frequency Distributions in Slavic Languages

In our endeavor to find an adequate theoretical model, we have concentrated on discrete frequency distibution models, rather than on continuous curves – for methodological reasons, which need not be discussed here. In order to test the goodness of fit of the models tested, we have employed $\chi^2$ tests. This traditional procedure is problematic, however, since the $\chi^2$ value linearly increases with sample size, the $\chi^2$ value thus becoming sooner significant – and in case of grapheme studies, we are almost always concerned with large samples. Therefore, we have relativized the latter by calculating the discrepancy coefficient $C = \chi^2/N$, considering a value of $C < 0.02$ to be a good, a value of $C < 0.01$ a very good fitting.

As to the models tested, we did not expect that one and the same model would be universally relevant, i.e. would be able to cover all languages of the world. We did not even assume that one model would be sufficient to cover all those (Slavic) languages which were the objective of our study. Therefore we have tested all those models which have been favored as successful rank frequency models in the past. Specifically, we tested the following distribution models (for details, cf. the studies mentioned above):

1. Zipf (zeta) distribution;
2. Zipf-Mandelbrot distribution;
3. geometric distribution;
4. Good distribution;
5. Whitworth distribution;
6. negative hypergeometric distribution.

It would be beyond the scope of the present paper to discuss the mathematical details of these distribution models, or the theoretical interrelations between them (cf. Grzybek, Kelih & Altmann 2004). Rather, it should be sufficient to summarize that for all three languages mentioned above, we found that the organization of the grapheme frequencies followed none of the traditionally discussed models. Rather, it was the negative hypergeometric distribution ($NHG$) – and only this model[1] – which turned out to be adequate; quite unexpectedly, all

[1] It should be noted that the allegedly exclusive validity of the $NHG$ distribution as a theoretical model claimed here relates only to the data we have analyzed thus far.

other models did not fulfill the above-mentioned criteria and thus had to be ruled out as adequate models.[2] Therefore, the $NHG$ distribution should briefly be presented here. It may be derived in different ways; here, it may be suffice to interpret it with recourse to Wimmer & Altmann's (2005a,b) *Unified Derivation of Some Linguistic Laws*, namely, in the form of equation (2):

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

Inserting in (2)

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

one obtains equation (3):

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

from which the $NHG$ results (with $x = 0,1,\ldots n$, $K > M > 0$, and $n \in \{1,2,\ldots\}$), as given in equation (4):

$$P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}} \tag{4}$$

Since in case of rank frequency distribution, the first class is $x = 1$, the $NHG$ has to be used in its 1-displaced form, as displayed in equation (5), with $x = 1,2,\ldots n+1$, $K > M > 0$, and $n \in \{1,2,\ldots\}$),

$$P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}} \tag{5}$$

This does not principally rule out all other models as possibly being relevant, and this is not to be misunderstood as a claim for a single universal model. Rather, there may be transitions between various model, or covergencies between them, and it is a matter of boundary conditions to be controlled in each single study, if one of the above-mentioned model, or eventually even other models not mentioned here, are more adequate.

[2] Only in case of Russian, the Whitworth distribution which, under particular conditions, is a special case of the $NHG$ (in its partial sums form), turned out to be an adequate model, too.

## 3 Three Case Studies: Russian, Slovene, Slovak

Thus far, the results of four case studies have been reported which were conducted to test the model described above. In the case study involving Russian (Grzybek, Kelih, & Altmann 2004), 37 samples composed of different genres were analyzed. The text corpus included literary texts by A.S. Puškin, L.N. Tolstoj, F.M. Dostoevskij, and A.P. Čechov, as well as a number of scientific texts. In order to control the factor of text homogeneity, all texts were individually analyzed as homogeneous texts. Additionally, text segments, mixtures, and cumulations were artificially formed on the basis of these texts and analyzed in this form, as well. Finally, they were put together and to build a complete corpus of ca. 8.7 million graphemes and analyzed as such.

As a result, the $NHG$ distribution turned out to be an adequate model for all 37 samples, with a discrepancy coefficient of $C < 0.02$ for each of them. Figure 1 illustrates the result for the complete corpus, where fitting the $NHG$ distribution resulted in a discrepancy coefficient value of $C = 0.0043$.

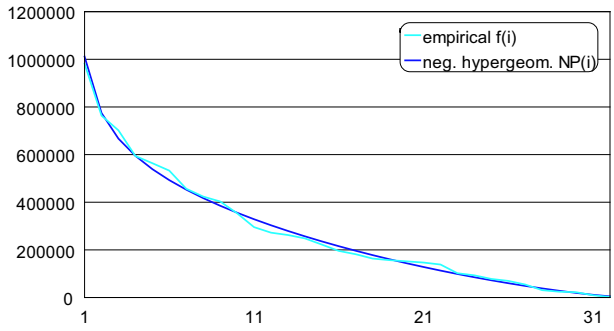



**Fig. 1.** Fitting the $NHG$ Distribution to Russian Corpus Data

In the Russian study, a first interesting observation was made as to the parameters of the $NHG$ distribution: Apart from parameter $n$ – which, with $n-1$, directly depends on the inventory size und thus is for all cases is constantly $n = 32 = 31$ in the case of Russian with its 32 different graphemes[3] –, also

[3] If one counts the Russian letter 'ë' as a separate letter, instead of realizing at as an allograph of the letter 'ë', the inventory size of the Russian alphabet increases to 33, of course. It is evident that, as soon as inventory size comes into play as an influencing parameter when fitting a given distribution to particular data, this question may turn out to be relevant for the results obtained. Therefore, in order to control this factor systematically, Grzybek, Kelih & Altmann (2006) have re-run their analysis of Russian material under three different conditions in thirty homogeneous texts: (a) texts in which the Russian letter 'ë' does not occur ($n = 32$), (b) texts containing the letter 'ë' ($n = 33$), and (c) the same texts as in (b), thus in principle containing the letter 'ë', but the latter a posteriori being transformed to 'e' ($n = 32$) for the

parameters $K$ and $M$ seemed to display a relative constancy across all samples (with $K \approx 3.16$ and $M \approx 0.82$), $K$ ranging from $2.95 \leq K \leq 3.42$, and $M$ ranging from $0.77 \leq M \leq 0.85$. Figure 2 illustrates the observed constancy of the results obtained, with $00.43 \leq C \leq 0.0169$.

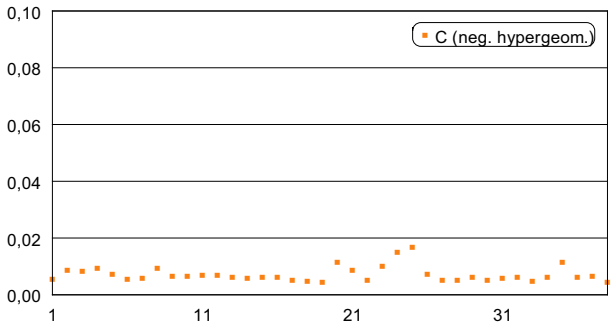



**Fig. 2.** $C$ Values for Fitting the $NHG$ distribution to Russian data)

Given these findings of the Russian case study, the idea was born to study the problem systematically for other Slavic alphabets, too. In this respect, Russian with its 32 (or 33) letters, has to be considered as having a medium inventory size as compared to other Slavic languages. Slovene, in turn, with its 25 letters, represents the minimum inventory size, and Slovak, with its 46 letters, is located at the upper end of the scale.[4]

In the Slovene study (Grzybek & Kelih 2003), twenty samples were analyzed, including literary texts and letters by Ivan Cankar, France Prešeren, Fran Levstik, as well as journalistic texts from the journal *Delo*; again, in addition to homogeneous texts, cumulations, segments and mixtures were artificially created and analyzed, as well as the complete corpus consisting of ca. 100.000 graphemes. As a result, the $NHG$ distribution turned out to be the only adequate model for all samples: the discrepancy coefficient was $C < 0.02$ in all cases (with $C = 0.0094$ for the corpus).[5]

Again, for the Slovene data, too, the values of the parameters $K$ and $M$ of the $NHG$ distribution turned out to be quite stable across all samples, with

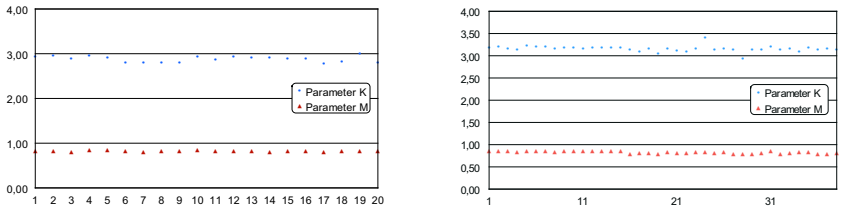
analytic purpose described above.– Since these data have not yet been published, the present article is based on the results reported in Grzybek, Kelih, & Altmann (2005).

[4] In case of Slovak, the inventory size decreases to 43, if one does not consider the digraphs 'ch', 'dz', and 'dž' to be separate letters in their own right.– Here, too, Grzybek, Kelih, & Altmann (2005a,b) conducted systemtatic studies to control the factor of defining the basic graphemic units.

[5] For Slovene, too, Grzybek, Kelih, & Altmann (2005) have re-run their analyses, extending the data basis to thirty homogeneous texts. As in case of Russian, the present study is based on the results reported by Grzybek & Kelih (2003).

$K \approx 2.89$ and $M \approx 0.81$), $K$ ranging from $2.79 \leq K \leq 3.01$, and $M$ ranging from $0.80 \leq M \leq 0.83$. Interestingly enough, no significant difference was observed between the group of homogeneous texts, on the one hand, and the artificially composed text samples (segments, cumulations, mixtures), on the other hand, as far as the parameter values of $K$ and $M$ are concerned (the mean values being $\bar{K} = 2.89$ and $\bar{M} = 0.81$, for both groups of texts as well as for all samples jointly). Thus, on the level of graphemic organizsation, text heterogeneity does not seem to play a crucial role.

A comparative inspection of Figure 3 shows that for each of the languages, parameters $K$ and $M$ are relatively constant, but that the constancy of parameter $K$ is realized on different levels, being slightly higher for Russian.




(a) Slovene Data ($n = 25$) (b) Russian Data ($n = 32$)

**Fig. 3.** Constancy of Parameter Values $K$ and $M$ ($NHG$ distribution)

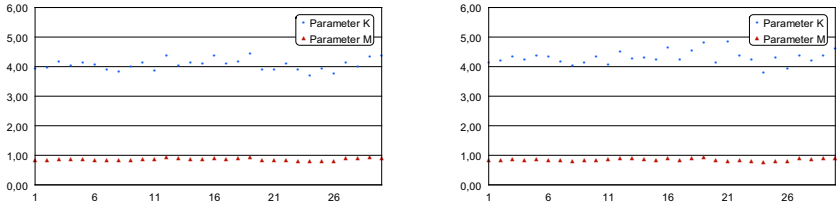
Given this observation, the hypothesis brought forth that not only parameter $n$ of the $NHG$ distribution, but also parameter $K$ might be particular function of the inventory size. In this case, the analysis of Slovak data, should yield additional arguments in favor of this assumption. Consequently, two studies were conducted, based on thirty Slovak texts, summing up to a corpus of ca. 150.000 letters. In the first of these two studies (Grzybek, Kelih & Altmann 2005a), Slovak grapheme frequencies were analyzed without taking into consideration the above-mentioned digraphs, the inventory size thus being $n = 43$; in the second study (Grzybek, Kelih & Altmann 2005b), the same material was analyzed, this time counting digraphs as a category in its own right, the inventory size thus rising up to $n = 46$.

As a result, the $NHG$ distribution once again turned out to be the only adequate model, under both conditions, with $K$ and $M$ displaying a relative constancy in either case. In case of the first study (with $n = 43$), the discrepancy coefficient was $C < 0.02$ in 28 of all 30 samples (with $C < 0.01$ in ten of the samples, and $C = 0.0102$ for the whole corpus); as to an interpetation of the finding that no good fitting was obtained for two of the samples, the authors referred to the fact that these two samples were extremely small with $N = 562$,

and $N = 446$ graphemes, respectively. Once again, the values of the parameters $K$ and $M$ of the $NHG$ distribution were relatively constant across all samples, with $K \approx 4.07$ and $M \approx 0.85$), $K$ ranging from $4.46 \leq K \leq 3.69$, and $M$ ranging from $0.78 \leq M \leq 0.94$.

In case of the second study (with $n = 46$), the results were slightly worse, with a discrepancy coefficient of $C < 0.02$ in 25 of all 30 samples (with $C < 0.01$ in five of the samples, and $C = 0.0139$ for the whole corpus). Yet, with $K \approx 4.31$ and $M \approx 0.84$), $K$ ranging from $4.86 \leq K \leq 3.81$, and $M$ ranging from $0.76 \leq M \leq 0.92$.

Figure 4 illustrates the observed constancies of parameters $K$ and $M$ for both conditions.




(a) Slovak ($n = 43$))

(b) Slovak ($n = 46$)

**Fig. 4.** Constancy of Parameters $K$ and $M$ ($NHG$ distribution; Slovak data)

By way of a preliminary summary, one can thus say that the two Slovak studies yield two important reults: first, the $K$ values of the first study (with $n = 43$), is indeed lower as compared to those of the second study (with $n = 46$); and secondly, the Slovak $K$ values, taken on the whole, are clearly higher as compared to those from the Slovene (with $n = 25$) and Russian (with $n = 32$) studies.

## 4 Consequences of the Single Case Studies

The four case studies reported above thus not only corroborated the initial hypothesis that the grapheme systems of the languages under study are systematically organized. Additionally, the findings clearly showed that the grapheme frequencies can be modelled with recourse to one and the same frequency distribution, namely, the $NHG$ distribution. Furthermore, the results obtained gave rise to further hypotheses as to a possible interpretation of at least one of the parameters of this model, namely, parameter $K$.

Taking into account the results for each language separately, it first seemed that the two parameters $K$ and $M$ are both relatively constant within a given

language. However, as soon as data for all three languages were available, it could be seen that parameter $K$ is definitely higher for a language with a larger inventory size, parameter $M$ not displaying such a direct increase. Grzybek, Kelih, & Altmann (2005a) therefore assumed this to be a hint at the possible (direct or indirect) dependence of parameter $K$ on inventory size, whereas parameter $M$ still seemed to be relatively constant across languages. The assumption of a direct dependence of $K$ on inventory size was therefore directly tested in Grzybek, Kelih, & Altmann's (2005a,b) studies on Slovak: For the sake of simplicity, they considered parameters $K$ and $M$ to be random variables with finite mean values and finite variances, and then compared the mean values of the parameters for both Slovak conditions ($n = 43$ vs. $n = 46$) by way of a $t$-test. As the results showed, parameter $K$ is significantly higher for $n = 46$ as compared to $n = 43$ ($t_{FG=56} = 4.53$; $p < 0.001$). However, a comparison of the mean values of parameter $M$ by way of a $t$-test showed that in this case, for both conditions ($n = 43$ vs. $n = 46$), there is no significant difference ($t_{FG=58} = 1.07$; $p = 0.29$).

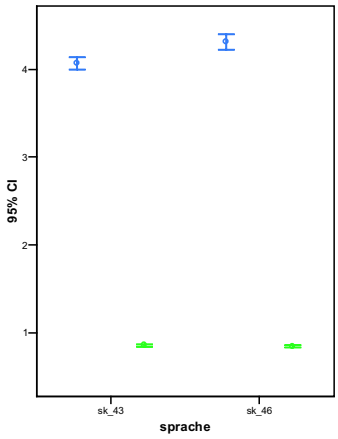
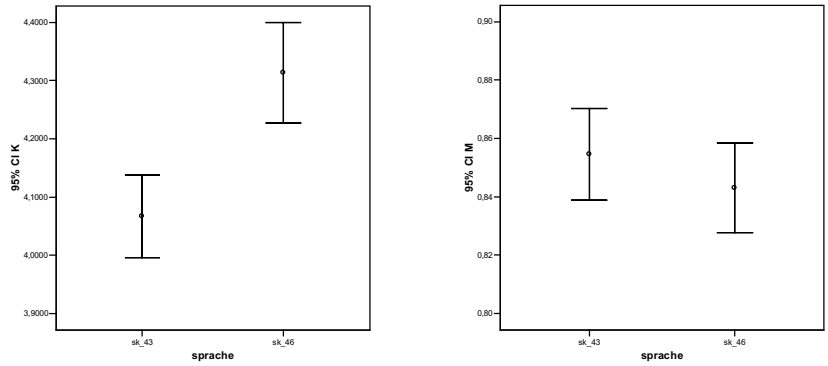



**Fig. 5.** Mean Values and Confidence Intervals for $K$ and $M$ (Slovak Data)

Fig. 5 illustrates the tendencies of both parameter values in form of a 95% confidence interval within which the relevant parameter may be expected with a 95% probability. It can easily be seen that parameter $K$ clearly differs for both conditions ($n = 43$ vs. $n = 46$), whereas parameter $M$ does not seem to vary significantly. The detailed Figure 6 additionally shows that the confidence intervals of $K$ do not overlap, whereas they do for parameter $M$.

Whereas there is thus some evidence that parameter $K$ may be directly related to inventory size, there is no such evidence with regard to parameter $M$. However, in their second study on Slovak graphemes, Grzybek, Kelih, & Altmann (2005b) found some other evidence of utmost importance, hinting at a direct relation between the two parameters, within a given language: under this




(a) Slovak ($n = 43$) (b) Slovak ($n = 46$)

**Fig. 6.** 95% Confidence Intervals for Parameters $K$ and $M$ (Slovak Data)

condition (i.e., with $n = 46$), they found a highly significant correlation between $K$ and $M$ ($r = 0.59, p = 0.001$). In a re-analysis of the Slovak data with $n = 43$, the very same tendency could be found, the correlation even being more clearly expressed ($r = 0.83$, $p < 0.001$).

The interpretation arising thus is that one of the two parameters ($K$) is dependent on inventory size (and thus particularly relevant across languages), whereas the second parameter ($M$) is relevant within a given language. As Grzybek, Kelih, & Altmann (2005b) state, we are concerned here with a highly promising perspective: if the findings obtained could be corroborated on a broader basis, an interpretation of both parameters $K$ and $M$ would be at hand.

This assumption needs further testing, of course, and the present study is, as was said above, a very first step in this direction. As was said above, it would be too daring to utter far-reaching conclusions at this time, and if so, only with utmost caution. The four case studies reported above do no allow for solid generalizations; first, they imply some methodological problems, and second, the number of languages is too small for any extrapolation of the results obtained. Yet, the impression arises that not only the grapheme frequencies of each language per se are systematically organized, but also, in addition to this, the organization of the graphemic systems in general. One argument supporting this assumption is the fact that the grapheme frequenices of all three languages studied follow one and the same model; this is only a minor argument, however, since a model may well be a special case of a more general one, or it may converge to a related model. A major argument in favor of the assumption brought forth, then, is the possible interpretation of the parameters.

Yet, there seems to be sufficient evidence to generalize the results obtained in form of the derivation of some working hypotheses for future research.

## 5 From Case Studies to Systematic Research: Towards a Theory of Grapheme Frequencies

A first step in the direction outlined might thus be a comparative analysis of the four studies reported above. Table 1 presents the results obtained in a summarizing manner.

**Table 1.** Mean Parameter Values and Confidence Intervals

| Language | $n$ | Parameter $K$ | | | Parameter $M$ | | |
|---|---|---|---|---|---|---|---|
| | | $\bar{K}$ | $K_\uparrow$ | $K_\downarrow$ | $\bar{M}$ | $M_\uparrow$ | $M_\downarrow$ |
| Slovene | 25 | 2.89 | 2.86 | 2.92 | 0.8115 | 0.8062 | 0.8168 |
| Russian | 32 | 3.16 | 3.14 | 3.19 | 0.8186 | 0.8105 | 0.8267 |
| Slovak | 43 | 4.07 | 4.00 | 4.14 | 0.8546 | 0.8389 | 0.8703 |
| Slovak | 46 | 4.31 | 4.23 | 4.40 | 0.8430 | 0.8276 | 0.8584 |

As a closer inspection of Table 1 shows, there seems to be a clear increase of parameter $K$ with an increase of inventory size ($n$), whereas parameter $M$ does not display a corresponding tendency; rather, parameter $M$ seems to be rather constant across languages. Fig. 7 illustrates these two tendencies.

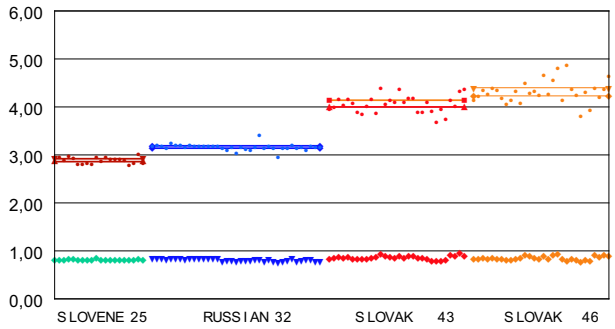
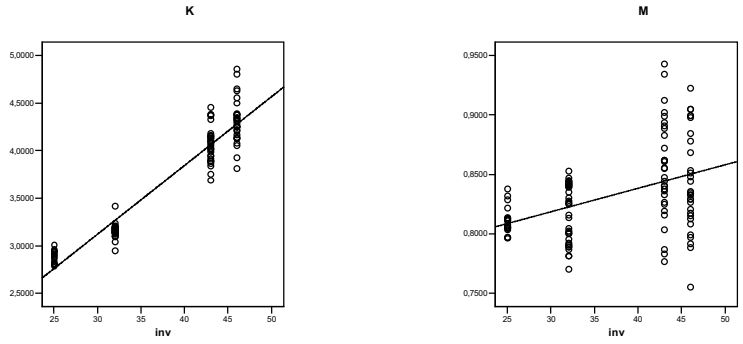



**Fig. 7.** Parameters $K$ and $M$ (With Confidence Interval) For Four Slavic Languages

Yet, as a statistical analysis shows, facts are more complex than it seems at first sight: Thus, calculating a bivariate correlation between the inventory size and the parameter values for $K$ and $M$, results in a correlation coefficient of $r = 0.956$ (for $K$) and $r = 0.424$ (for $M$), both correlations being highly significant ($p < 0.001$), the correlation for $K$ being more clearly expressed as compared to $M$. Figure 8 displays the result of regression analyses with inventory size as independent variable, $K$ and $M$, respectively, as dependent variables.




(a) Parameter $K$ (b) Parameter $M$

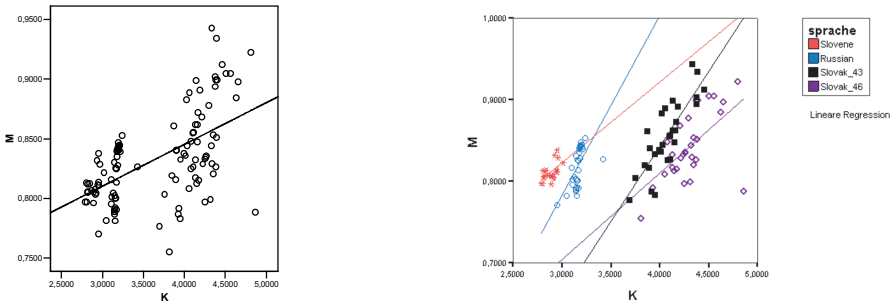
**Fig. 8.** Dependence of Parameters $K$ and $M$ on Inventory Size

The impression arising thus is that both $K$ and $M$ might depend on inventory size; this interpretation is weakened, however, or specified, by a closer analysis of the relation between both parameters. Given the finding that the correlation between parameter $K$ and inventory size is expressed more clearly (see above), it seems reasonable to take into consideration the possibility that $M$ is only indirectly dependent on inventory size, and directly on $K$. In fact, the correlation between $K$ and $M$ is highly significant ($r = 0.57$, $p < 0.001$). Figure 9 illustrates this tendendy; as a closer inspection shows, however, the dependence seems to be much more clearly expressed not across languages, but within a given language.

This observation may then be interpreted in terms of a direct (linear) dependence of parameter $K$ on inventory size $n$, and a direct (linear) dependence of parameter $M$ on parameter $K$. Consequently, parameter $M$ may be interpreted in terms of an indirect dependence on $n$. At this point, two perspectives emerge as possible orientations for future studies:

1. The first perspective is directed toward the study across languages; if in this respect, inventory size ($n$) is directly relevant for $K$, then it seems reasonable to concentrate on the mean values of $K$ for each language ($\bar{K}$).
2. The second perspective concentrates on processes within a given language; if $M$ indeed depends rather on $K$, within a given language, and less on $n$, then $K$ must be studied for each language individually ($K_i$).

As was shown above, $\bar{K}$ seems to be a linear function of $n$, thus being characterized by the equation $\bar{K} = h(N) = u \cdot N + v$. Furthermore, it now turns out that in fact $M_i$ seems to be a linear function of $K_i$, within a given language,




(a) Across Languages (b) Language-specific

**Fig. 9.** Dependence of Parameter $M$ on $M$

characterized by the linear function $M_i = a_i \cdot K_i$. Applying this formula to the data described above, one obtains the values represented in Table 2.

**Table 2.** Linear Dependences Between $K$ and $M$

| Language | $n$ | $\bar{K}$ | $\bar{M}$ | $a$ |
|---|---|---|---|---|
| Slovene | 25 | 2.8874 | .8115 | .280948 |
| Russian | 32 | 3.1636 | .8186 | .258741 |
| Slovak | 43 | 4.0666 | .8546 | .210090 |
| Slovak | 46 | 4.3137 | .8430 | .195181 |

As a closer inspection of Table 2 shows, we are not yet at the end of our interpretations: quite obviously, $a_i$ stands in a direct (linear) relation with $n$, which may be expressed by way of the formula $a_i = f(N) = c \cdot N + d$, the regression being almost perfect with a determination coefficient of $R^2 = .99$.

The observed tendency is illustratively presented in Figure 10), from which the linear decline of $a$ with increasing inventory size becomes evident.

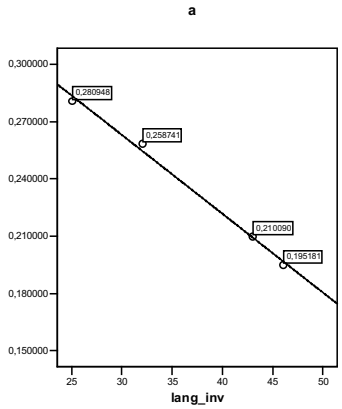



**Fig. 10.** $a$ and $n$

## 6 Perspectives

It goes without saying, and it has been emphasized repeatedly, that at this moment these interpretations are rather daring. More material, and more systematically chosen material, must be analyzed to put our assumption on a more solid ground. Still, if additional evidence can be gathered for the plausible interpretations outlined above, a scheme as depicted in Table 3 might be derived to describe this situation.

**Table 3.** A General Schema of Dependences

| | |
|---|---|
| $\bar{K} =$ | $h(N) = u \cdot N + v$ |
| $M_i =$ | $g(K_i) = a_i \cdot K_i$ |
| $a_i =$ | $f(N) = c \cdot N + d$ |

If the assumptions and hypotheses outlined above would indeed receive further support, we were in a lucky situation, which is highly desirable in quantitative linguistics, since we would be able to interpret all parameters of the theoretical distribution and thus have a qualitative interpretation. If the hypothesis brought forth above can be corroborated on a broader and more solid basis, including further (Slavic) languages, this might be relevant not only for linguistics. Ultimately, this would be a highly tricky mechanism from a broader perspective as well, relevent for systems theory and synergetics, in general: from this point of view, we are concerned with a low-level system of units relevant for the formation of higher-level units; on this low level the system's behavior is

determined merely by the inventory size of the units involved, and any variation on this level would be "corrected" by a second parameter, thus guaranteeing the system's flexible stability.

Only thorough research can show if our assumptions stand further empirical testing – the fate of science, though...

## References


1. Altmann, G.: Status und Ziele der quantitativen Sprachwissenschaft. In Jäger, S., ed.: Linguistik und Statistik. Vieweg, Braunschweig (1972) 1–9
2. Altmann, G.: Mathematische Linguistik. In Koch, W., ed.: Perspektiven der Linguistik. Kröner, Stuttgart (1973) 208–232
3. Grzybek, P., Kelih, E.: Grapheme Frequencies in Slovene – a Pilot Study. In Benko, V., ed.: Slovko 2003, Bratislava (2003) (to appear)
4. Grzybek, P., Kelih, E.: Grapheme Frequencies in Slovene. Glottometrics **12** (2006) (to appear)
5. Grzybek, P., Kelih, E.: Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens. Glottometrics **9** (2005) 62–73
6. Grzybek, P., Kelih, E.: Graphemhäufigkeiten (Am Beispiel des Russischen). Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. Anzeiger für slawische Philologie **31** (2003) 131–162
7. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten im Slowakischen (Teil I: Ohne Digraphen). In Nemcová, E., ed.: Philologia actualis slovaca. UCM, Trnava (2005) (to appear)
8. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten (Am Beispiel des Russischen). teil III: Systematische Verallgemeinerungen. Anzeiger für slawische Philologie **33** (2005) (to appear)
9. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten im Slowakischen (Teil II: Mit digraphen). In: Sprache und Sprachen in Mitteleuropa. GeSuS, Trnava (2005) (to appear)
10. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung. Anzeiger für slawische Philologie **32** (2004) 25–54
11. Wimmer, G., Altmann, G.: Towards a Unified Derivation of Some Linguistic Laws. In Grzybek, P., ed.: Contributions to the Science of Language: Word Length Studies and Related Issues. Kluwer, Dordrecht (2005) (to appear)
12. Wimmer, G., Altmann, G.: Unified Derivation of Some Linguistic Laws. In Köhler, R., Altmann, G., Piotrowski, R.G., eds.: Handbook of Quantitative Linguistics. de Gruyter, Berlin (2005) (to appear)

# Computer Treatment of Slavic and East European Languages

Third International Seminar
Bratislava, Slovakia, 10–12 November 2005
Proceedings

Editor
Radovan Garabík

Reviewers
Peter Ďurčo
Jana Levická

VEDA
Vydavatel'stvo
Slovenskej akadémie vied
Bratislava 2005