[Unpublished manuscript; submitted to Benko, V. (eds) (2003): Slovko 2003, 7 p.]

Peter Grzybek / Emmerich Kelih (Graz, Austria)

Grapheme Frequencies in Slovene

The present study focuses on grapheme frequencies in Slovene. It goes without saying, that counting letters (or graphemes), presenting the corresponding absolute (or relative) frequencies in tables, or illustrating the results obtained in figures, is only one particular step in a properly designed quantitative study. More often than not, however, it has rather been the last step, researchers being content with results being presented in the above-mentioned manner, and considering this to be the ultimate objective of their research. Yet, strictly speaking, providing the data is not even the first step if one understands quantitative linguistics to be a theoretical discipline, and not a data-heaping hobby.

In a more demanding perspective, 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).

In order not to be misunderstood, then: providing and presenting data is no useless or senseless waste of time; rather, it is part of scientific research, and it is a necessary pre-condition for theoretical models to be 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 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 over samples; the focus will thus be on the analysis of an underlying frequency distribution model.

In following the second of these two courses in this study, then, our hypothesis is that the frequency with which graphemes in a text occur, is not by chance, but regulated by particular rules; more specifically, our hypothesis says that this rule works in relative independence of the data structure (i.e., with 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 the function $P_x \sim$

$g(x)P_{x-1}$; more specifically, we expect the organization of the grapheme frequencies to follow not any one of the traditionally discussed models (zeta, Zipf-Mandelbrot, geometric, Good distributions), but either the so-called Whitworth distribution or the negative hypergeometric distribution (if not both of them), since these two models have been proven to be adequate for other Slavic languages, too (cf. Grzybek 2004, Grzybek, Kelih & Altmann 2004). 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, here (cf. Grzybek, Kelih & Altmann 2004).

In our empirical testing of these hypotheses, we will pay particular attention to the important factor of data homogeneity, analyzing complete texts (of different complexity) as well as text segments, mixtures, cumulations, and corpora. Given the results to be obtained, and given their statistical interpretation it may still be too early to arrive at a linguistic interpretation of the results, since not only more data from more (Slavic) languages are necessary, but also comparisons to their phonemic structure(s).

Within this general framework, the following data sets have been analyzed:

Table 1: List of Authors and Texts

No.	Author	Text	Chapter	Abbrev.	N
1	Cankar, Ivan	Greh	1	CI-G1	12301
2		Greh	2	CI-G2	11477
3		Greh	3	CI-G3	11685
4	Cankar, Ivan	Brief 1		CI-B	3971
5		Brief 2		CI-B	5191
6	Prešeren, France	Krst pri Savici		PF-K	11390
7	Levstik, Fran	Martin Krpan z Vrha		LF-M	22976
8		Pokljuk		LF-P	13581
9	Delo	Virus resnice (comment)		J-1	4237
10	Delo	Odštevanje (comment)		J-2	1900
11	*Complete corpus*			GK	**98709**

Table 2: List of Text Cumulations, Segments, and Mixtures

No.	Author	Text	Chapter	Abbrev.	N
12	Cankar, Ivan	Greh	1-2	CI-G(1-2)	23778
13			1-3	CI-G(1-3)	35463
14	Cankar, Ivan	Greh	1+3	CI-G1 + CI-G3	23986
15	Cankar Ivan & Delo	Greh & text 9		CI-G(1-3) + J-1	39700
16	Cankar Ivan & Prešeren France	Greh & Krst pri Savici		CI-G(1-3) + PF-K	46853
17	Prešeren, France & Levstik, Fran	Krst pri Savici & Martin Krpan z Vrha		PF-K + LF-M	34366
18	Levstik, Fran & Delo	Pokljuk & text 10		LF-P + J-2	15481
19	Cankar Ivan	Greh	every 4th line	CI-G(1/4)	9464

20	Prešeren France	Krst pri Savici	every 5th line	PF-K(1/5)	2317

Neither the distribution models nor the results obtained can be discussed in detail, here (cf. Grzybek, Kelih & Altmann 2004). Therefore, fig. 1 presents the most important findings in a synoptical form, showing for all data sets the value of the discrepancy coefficient $C = X^2 / N$, which for $C < 0.02$ indicates a good, for $C < 0.02$ a very good fitting result.

The six distributions listed above have been tested for their adequacy to model grapheme frequencies in Slovene. The results cannot be presented in detail, here (cf. Kelih & Grzybek 2004a); therefore it may be sufficient to provide a visual impression of the overall results. Fig. 1a illustrates the *C* values for the zeta and Zipf-Mandelbrot distributions, fig. 1b for the right truncated geometric and the right truncated Good1 distributions.

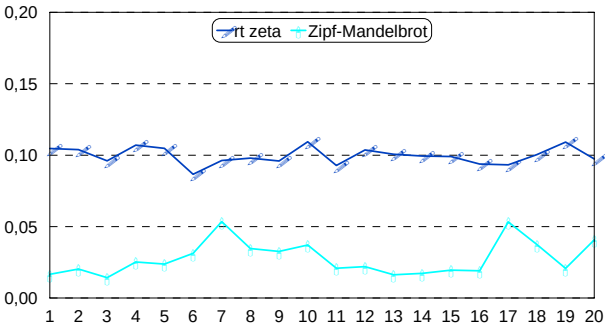

Fig 1a: Discrepancy coefficients *C* for zeta and Zipf-Mandelbrot distributions

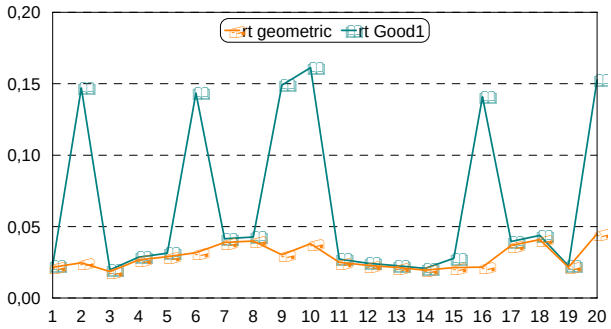

Fig 1b: Discrepancy coefficients *C* for right-trunc. geometric and Good1 distributions

As can be seen, none of these four distribution models yields satisfying results: for the zeta, the right truncated geometric, and the right truncated Good1 distributions, the *C* value is $C < 1$ in maximally two cases; the Zipf-Mandelbrot distribution is not much better with $C < 0.02$ in only six cases, as well.

This overall result corresponds to our prediction, which is based on our experience with the grapheme systems of other Slavic languages. As to the above-mentioned specific hypothesis, saying that either the Whitworth, or the 1-displaced negative hypergeometric distributions, or both of them may be better suitable, let us first present a visual impression, as well (cf. fig. 2).

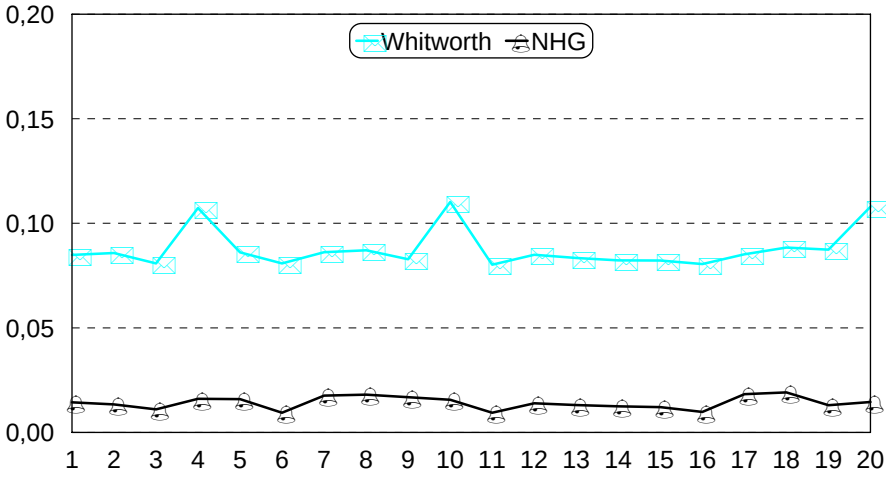

Fig. 2: Discrepancy coefficients *C* for Whitworth and neg. hypergeometric distributions

It becomes clear from fig. 2, that as opposed to Russian, the Whitworth distribution is no acceptable model for the ranked frequencies of Slovene letters, none of the *C* values being $C < 0.02$. The only model which adequately models the ranked frequencies of the 25 Slovenian graphemes in all samples, is the (1-displaced) negative hypergeometric distribution:

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

$$K > M > 0;\ n \in \{1,2,\ldots\}$$

Table 2 / fig. 1 present the results for the complete corpus.

Table 3: Negative hypergeometric distribution (corpus)

i	f(i)	NP(i)	i	f(i)	NP(i)
1	11305	12309,30	**14**	3255	2972,09
2	10107	9516,29	**15**	2972	2673,92
3	9149	8220,61	**16**	2342	2386,95
4	8599	7339,23	**17**	1979	2110,08
5	6344	6650,33	**18**	1768	1842,43
6	5676	6072,57	**19**	1606	1583,33
7	4954	5567,34	**20**	1462	1332,29
8	4670	5113,37	**21**	1046	1088,97
9	4360	4697,70	**22**	885	853,24
10	4094	4311,87	**23**	710	625,22
11	3832	3950,06	**24**	573	405,45
12	3552	3608,09	**25**	39	195,40
13	3430	3282,88			
K =	2,8654		χ^2 =	925,90	
M =	0,8072		FG =	21	
n =	24		C =	0,0094	

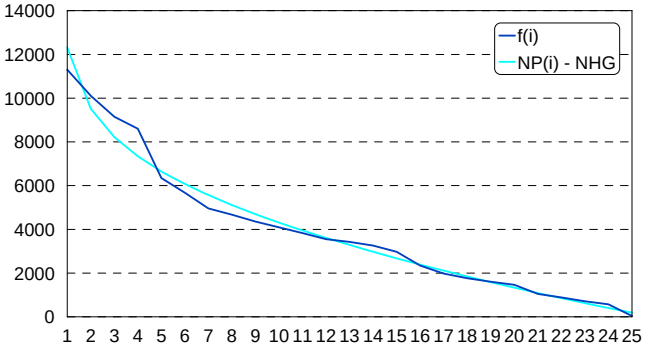

Fig. 2
Fitting the negative hypergeometric distribution (complete corpus)

A comparison with the results for all individual sample confirms the finding that the negative hypergeometric distribution is an adequate model for ranked grapheme frequencies in Slovene: in all cases, the discrepancy coefficient is C < 0.02, in three cases even C < 0.01. The results for all samples are represented in table 4, which also contains the values of the parameters *K* and *M* for each sample, as well as the chi and *C* values.

Table 4: Parameters of the negative hypergeometric distribution

No.	Text	K	M	No.	Text	K	M
1	CI-G1	2,9455	0,8138	11	*complete corpus*	2,8654	0,8072
2	CI-G2	2,9524	0,8114	12	CI-G(1-2)	2,9437	0,8120
3	CI-G3	2,8875	0,7966	13	CI-G(1-3)	2,9174	0,8051
4	CI-B	2,9631	0,8289	14	CI-G 1 + CI-G3	2,9092	0,8036
5	CI-B	2,9290	0,8321	15	CI-G(1-3) + J-1	2,9059	0,8082
6	PF-K	2,8141	0,8058	16	CI-G(1-3) + PF-K	2,8867	0,8057
7	LF-M	2,8045	0,7972	17	PF-K + LF-M	2,7884	0,7976
8	LF-P	2,8171	0,8053	18	LF-P + J-2	2,8362	0,8130
9	J-1	2,8028	0,8134	19	CI-G(1-4)	3,0146	0,8221
10	J-2	2,9484	0,8383	20	PF-K(1-5)	2,8170	0,8123

It could already be seen from fig. 2, that the discrepancy coefficient *C* turns out to be convincingly stable over all samples. Therefore, it is particularly interesting to note that also the parameters of the negative hypergeometric distribution display a convincing degree of stability: In addition to the fact that parameter *n* (which is defined by the inventory size minus one and therefore is constantly *n* = 24) remains unchanged, of course, also the values for *K* and *M* are extremely stable: $3.01 \geq K \geq 2.79$ und $0.84 \geq M \geq 0.80$. Figs. 3 illustrates the constancy of these results.

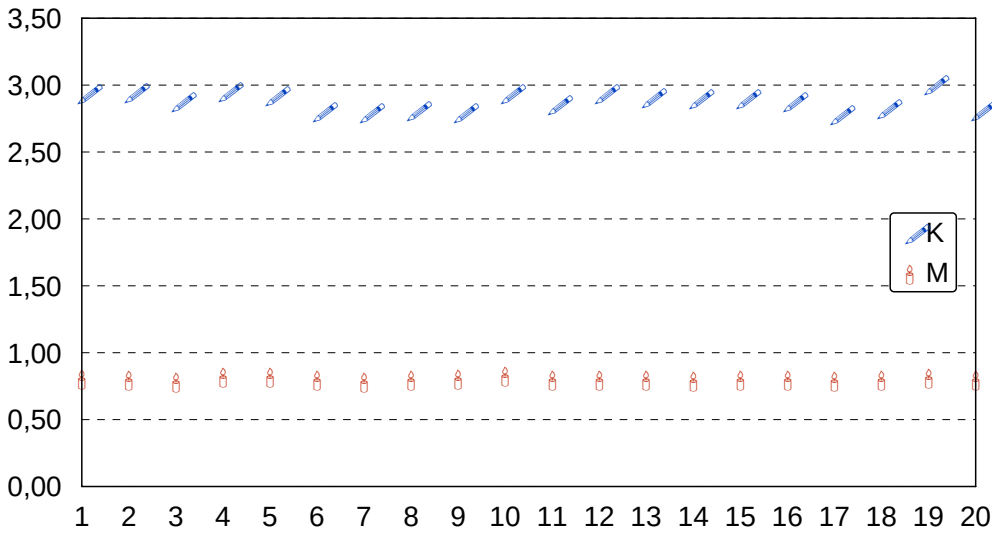

Fig. 3: Constancy of the parameters *K* und *M*

These results support the previously uttered hypothesis that there might be a chance to arrive at a qualitative (linguistic) interpretation of the results obtained, in future. At the present status of knowledge, a number of less daring conclusions may be more adequate:

1. The Slovene grapheme system turns out to be an orderly organized system, as far as the frequency of its elements are concerned.
2. The question of data homogeneity obviously plays only a minor, if any role on the level of grapheme analyses; the results are relatively constant, as long as the theoretical model is adequate, irrespective of the fact, if texts, text segments, text mixtures, or text cumulations are analyzed.
3. It is an important finding that five distributions, which have been assumed to be adequate models in previous research (zeta, Zipf-Mandelbrot, geometric, Good1, Whitworth) are not acceptable for Slovene grapheme frequencies; most likely, a number of assumptions about other languages will have to be modified.
4. The negative hypergeometric distribution turns out to be an adequate model, what seems to be the case in many other Slavic languages, too. A problem thus far unsolved is the fact that only one of its parameters (*n*) allows for an easy interpretation. Yet, the constancy of the other two parameters (*K* and *M*) allow for the hypothesis that there might me some qualitative explanation of the empirical findings. The extension of the studies to further Slavic languages may provide relevant insights.
5. In further pursuing this line of research, it will be of utmost importance to search for cross-references with the corresponding phonemic systems (cf. Grzybek & Kelih 2004b). As a result, answers will be obtained with regard to the question, in how far the models discussed here are particularly (or exclusively?) adequate for Slavic languages, which display a relative great (though diverging) proximity to the corresponding phoneme structures.

References

Altmann, G. (1972): „Status und Ziele der quantitativen Sprachwissenschaft.“ In: Jäger, S. (ed.), *Linguistik und Statisti*k. Braunschweig. (1-9).

Altmann, G. (1973): „Mathematische Linguistik.“ In: Koch, W.A. (Hrsg.), *Perspektiven der Linguistik.* Stuttgart. (208-232).

Grzybek, P. (2004): “A Study on Russian Graphemes.” In: *Festschrift für T.M. Nikolaeva*. Moskva. [In print]

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“, in: *Anzeiger für slawische Philologie, 31*; 131-162.

Grzybek, P.; Kelih, E. (2004a): „Graphemhäufigkeiten im Slowenischen: Theorie und Empirie“. [In prep.]

Grzybek, P.; Kelih, E. (2004b): “Zusammenhänge zwischen Graphem- und Phonemhäufigkeiten (am Beispiel des Slowenischen).” [In prep.]

Grzybek, P.; Kelih, E.; Altmann, G. (2004): „Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung“, in: *Anzeiger für slawische Philologie, 32.* [In print].