Peter Grzybek (*Graz, Austria*)

A STUDY ON RUSSIAN GRAPHEMES *

Formal aspects of the Russian alphabet have repeatedly been the object of linguistic studies. These studies need not be enumerated in detail, here; it may suffice to refer to the collective monograph *Опыт описания русского языка в его письменной форме* (Moskva 1964), written by Z. M. Volockaja, T. N. Mološnaja, and T. M. Nikolaeva. This book was an important step in studying the Russian graphemic system as a sign system in its own right, and it must be seen in the context of further related studies, as those by Nikolaeva [1961a, b; 1965; 1969][1].

Ultimately, any study of the Russian graphemic system (as of other graphemic systems, as well, of course) will be faced with one basic question, namely, which elements which must be taken into consideration, representing the basic components of the system under study? For most alphabets, an answer to the basic question as to the size and the (number of) elements of the alphabetic system under study, depends on the decision if the alphabet in question is regarded to be an autonomous or a heteronomous system; in principle, this topic has been discussed in detail by T. M. Nikolaeva [1965: 130f.], though with a different terminology. Basically, the decision to regard the graphemic system of a given language as an autonomous or a heteronomus system, is identical with an answer to the question if a grapheme (or a letter) is considered to be a (linguistic) sign in its own right, or if it is regarded as the constituent of a sign, only. Considering a writing system to be an autonomous sign system, graphemes are no signs since they have no meaning, but only have differentiating function; as opposed to this, according to an heteronomous perspective, each grapheme has sign character, since it is regarded to be the sign of a phoneme of the given language (and thus has the function of a secondary sign).

It is obvious that any decision as to autonomy or heteronomy of a writing system is of immediate relevance for the question which elements (and, of course, how many elements) are to be considered graphemes of the given writing system:

a. Whereas from a heteronomous perspective, those letters and letter combinations have to be defined as 'graphemes', which correspond to a phoneme of the language in qu-

* The present study was written in context of the research project «Word Length Frequency Distributions», financially supported by the Austrian Fund for Scientific Research (project P-15485) — as to the general background cf.: http://www-gewi.uni-graz.at/quanta

estion. Thus, in English, for example, not only the single letters *s* and *h*, but also the letter combination *sh* would have to be considered as graphemes, since *s*, *h*, and *sh* all signify individual phonemes; therefore, in English, the three graphemes *s*, *h*, and *sh* would be formed from the two letters *s* and *h*.

b. Following an autonomous definition, graphemes can be defined by way of commutation tests on the basis of written texts. Thus, replacing the letter *h* by *r* in the English word *thick* results in the word *trick*; both *h* and *r* would thus have to be considered autonomous graphemes in English, and *h* would not only be part of the complex grapheme *th*. Still, according to the autonomous point of view, as well, graphemes are not identical with letters: thus, in the Old English letters ≠ and Δ, for example, can be considered to be allographs of one grapheme, since they can replace each other in the Old English orthography. Furthermore, under particular conditions, also the distinction between small and capital letters may be relevant with regard to the (size of the) grapheme inventory, from an autonomous perspective: although any such pair corresponds to one phoneme only, different meanings of a word may result from the fact if a word is written with small or capital letters (cf. *fest* or *Fest*, in German).

In the above-mentioned study by Volockaja, Mološnaja, and Nikolaeva (1964), an attempt is undertaken to describe the inventory of Russian graphemes on the basis of their constitutive elements (i.e., of the "figures", in Hjelmslev's sense, of which a sign is construed). A ‚grapheme' is defined as an "an abstract unit of the alphabet, which may have for expression forms: printed or hand-written, small and capital, these four expression variants are called a l l o g r a p h s or l e t t e r s" (ibd., 10). In this understanding, then, a grapheme, being an abstract concept (as opposed to an allograph, or a letter), may not be drawn, or written. Based on this assumption, one obtains the 33 standard graphemes for Russian language, which are represented in table 1 (cf. ibd., 11).

Grapheme number	Grapheme name	Grapheme number	Grapheme name	Grapheme number	Grapheme name
1	а	12	ка	23	ха
2	бэ	13	эль	24	цэ
3	вэ	14	эм	25	чэ
4	гэ	15	эн	26	ша
5	дэ	16	о	27	ща
6	йэ	17	пэ	28	йэр
7	йо	18	эр	29	ы
8	жэ	19	эс	30	йэрь
9	зэ	20	тэ	31	э
10	и	21	у	32	йу
11	и краткое	22	эф	33	йа

Given this inventory of 33 graphemes, "each of the four distinguished types represents an independent system of interrelated signs" (ibd., 19). As a consequence, each of the four variants has to be studied separately. Concentrating on the capital letters in their printed form, the authors conclude that of all 33 graphemes, the majority coincides with their small equivalents, being different only in size.

А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
а	б	в	г	д	е	ё	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

As can be seen, the shape of 29 letters coincides, differences exist only for the letters *а*, *б*, *е*, *ё*. Of course, since differences are mainly to be observed on the level of letters, not of graphemes, the number of differences strongly depends on the concrete realization of the written or printed allograph[2]. Thus, in their attempt to describe the inventory of Russian graphemes in their printed form, the authors concentrate on capital letters. According to their opinion, four elements are considered to be relevant (and sufficient) for their systematic description:

1. a skew line: /
2. a horizontal line: —
3. a vertical line: |
4. a bow: ⊃

Figure 1 represents the generation of the individual graphemes from these four constitutive elements.

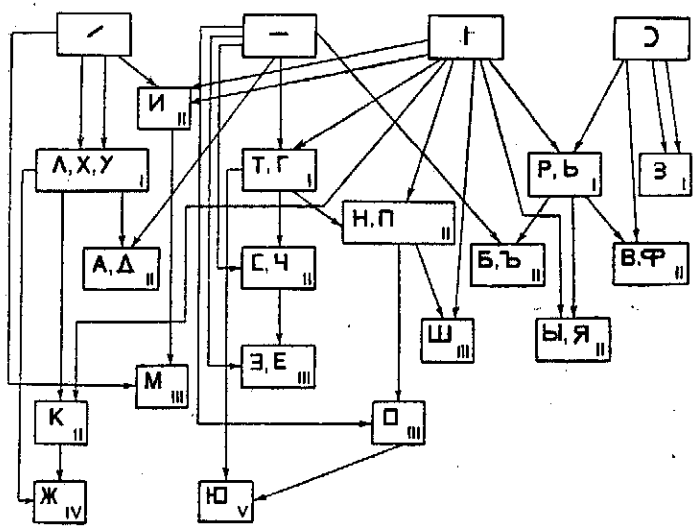


Fig. 1: Generation of capital printed letters

As can easily be seen from fig. 1, a number of graphemes are excluded from the analysis: **Ё, Й, Щ, Ц**. The reason for their exclusion is the fact that each of them contains one of the following diacritical elements: ˘ " ˥ .

According to the authors, these diacritical signs are not considered to be graphemes, in Russian, because they exist only along with one letter, each which, being relevant for the distinction of only one letter pair (Е-Ё, Й-И, Ш-Щ). In this respect, the letter **Ц** represents an exception since there is no corresponding letter without the diacritical element ˥ in the Russian alphabet. In the authors' opinion, the four graphemes pointed out above therefore represent separate graphemes (as opposed to Latin alphabets, where they are to be understood as the combination of a basic letter plus diacritical sign).

As can be seen from figures 1 and 2, there remain 29 graphemes which are analyzed with regard to their constitutive elements. One may agree with this reduction of the graphemic inventory, or not; still, the idea is intriguing that these letters represent different degrees of "complexity" (complexity being interpreted as the number of constituent elements per letter). In fact, as can be seen from fig. 2, four ranks can be distinguished, with an increase of one element per grapheme, the minimum of elements being two (rank one).

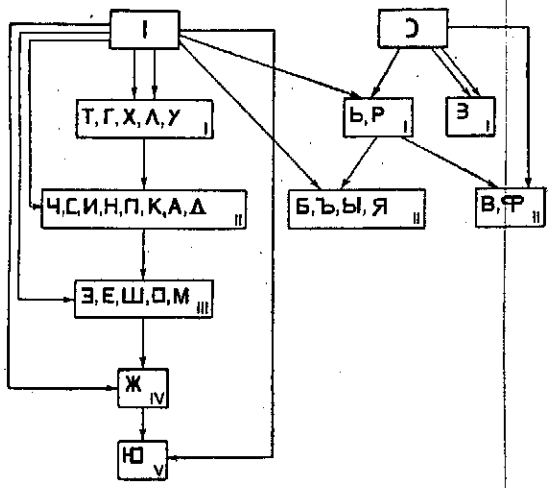
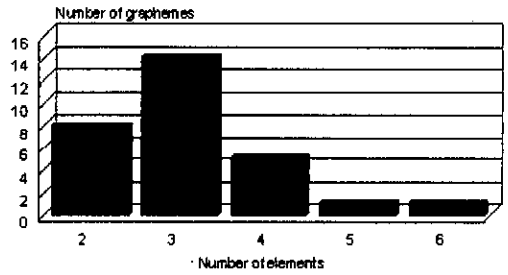


Fig. 2: Generation of ranks for capital printed letters

It can easily be calculated how many graphemes with a given number of elements are part of the alphabet. The corresponding data are represented in table 1. The graphic representation of these data in figure 3 clearly shows the left-skewed asymmetry of the frequency distribution ($\gamma = 1.286$), which significantly deviates from a normal distribution (with a Shapiro-Wilk test value of 0.811, $p < 0.001$).

This characteristic asymmetry coincides with observations from other linguistic units and levels, and it asks for more detailed analyses. In this respect, two *caveats* should be carefully taken into account, however:

number of elements	number of graphemes
2	8
3	14
4	5
5	1
6	1



Tab. 1 / Fig. 3: Frequencies of letter complexity

1. It seems that Nikolaeva and her co-authors, at the time of writing their book, were guided by the idea to provide a minimal inventory of distinctive features, thus striving for a maximum economy of meta-language. In fact, Nikolaeva [1969: 483] assumed that for the description of a given inventory M, the number of distinctive features should be $N = \log_2 M$. According to this calculation, we should thus arrive at a number of five elements, for the Russian alphabet (which is relatively close to the inventory used). However, concentrating on the economy of meta-language, is only one possible perspective. From a synergetic point of view, there are other needs of the graphemic system to be taken into consideration; in this respect, we are concerned not only with factors, such as minimization of coding or decoding effort, but also with the maximization of transfer safety (and thus with redundancy), with specification and distinctiveness, diversity, etc. From this perspective, many elements of Russian graphemes must not be considered to be "superfluous" (cf. [Volockaja et al. 1964: 25]), and it seems reasonable to follow the authors' suggestion to clearly distinguish between the level of abstract graphemes and sign elements of natural language. — As a consequence, more elements will be need for an adequate description of Russian letters as they occur in reality, which, in turn, will result not only in a change of the inventory (size), but, quite naturally, also of the frequency of the constituting elements[3].
2. Whereas Nikolaeva and her co-authors have concentrated on the level of the graphemic system, thus allowing for the frequency analysis of the given elements on the systemic level, it seems reasonable to ask a parallel question, too, which, thus far, has been largely neglected, at least with regard to standard European alphabets. This question may be phrased in the following way: how often do graphemes with a particular number of constituting elements occur in a running text, and what is the frequency of the elements themselves, under this condition? Irrespective of the fact that different classification systems will arrive at different data, this question is much less naive than it may sound at first sight. Ultimately, it will give an answer not only as to the state of the (alphabetic) system, its redundancy and discriminatory potential, but also as to basic processes of production and reception. It would not be economic, for example, if letters constituted of only a few elements, would occur rarely, and

letters with more elements would be of high frequency. In other words, it would be possible to gain insight into the efficiency of the graphemic system (and eventually of the system of its description), not only in a paradigmatic perspective.

The present study concentrates on one related question only, asking for the frequency of Russian graphemes. Yet, even in confining ourselves to this particular perspective, there remain two major directions of research. Given the frequency of Russian graphemes, based on a particular sample (be it a single text, part of a text, a mixtures of texts, or a corpus), 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 the history of the study of Russian grapheme frequencies, practically only the first course has been followed, although the question as to an overall theoretical model has been implicit in many of them (for a systematic historical overview and analysis of quantitative studies of Russian graphemes cf. [Grzybek/Kelih 2003a]).

In fact, in order to further analyze the componential structure of the Russian graphemic system, i.e., in studying the frequency of their constitutive elements, not only a comprehensive system for describing the totality of all 33 Russian graphemes with all their elements — which may vary from realization to realization — will be necessary; also, knowledge of each individual grapheme's frequency is a *sine qua non*. It goes without saying that such a new componential analysis cannot be developed here, "*en passant*". Therefore, we shall confine ourselves here to dealing with the second issue outlined above, presenting an attempt to describe a general frequency distribution model for Russian graphemes. We do not only consider this to be an important contribution to the "Description of Russian in Its Written Form"; we also consider this to be a proof of the systematic nature of the Russian grapheme system.

The question of an overall theoretical distribution model for graphemes has only rarely been asked explicitly (not only, as far as Russian graphemes are concerned); generally speaking, research along this line has been done, to name but the most important studies, by Sigurd [1968], Good [1969], Gusein [1988], or Martindale et al. [1996]. The overall interest of works like these was not so much the frequency of individual graphemes, but an answer to the question which relative frequency the most frequent, the second most frequent, the third most frequent, etc. graphemes have. The focus of these studies thus has been a so-called rank frequency distribution, the aim of theoretical modeling ultimately being a mathematical formalization of the distance(s) between the individual frequencies: Transforming given frequencies into a descending order, and graphically relating the data points with each other, the result is not a linear decline; rather, one obtains a specific, monotonously declining (usually hyperbolic) curve. And the idea is to model

the exact form of this curve in order to see, if the frequencies of different samples (i.e., the specific kind of decline) has one and the same shape.

However, all of the above-mentioned studies have a number of methodological flaws, which need not be discussed here in detail, but still should be mentioned *in toto*:

1. More often than not, the graphem(at)ic and phone(ma)tic levels of language are not consequently distinguished from each other, assuming that these two linguistic units (or forms of representation) follow one and the same model. This assumption seems to be reasonable, of course; still, it is more proper to clear keep apart these different levels of description, at least in a first approximation to the question.
2. Usually, research has not paid due attention to differences in the quality of the data material, provided on the basis of texts, parts of texts, text cumulations or mixtures (corpora), thus neglecting the important condition of data homogeneity. Again, it may well be that this factor is not relevant for the analysis of graphemes; still, the factor should carefully controlled, as in studies as well.
3. The elaboration of relevant frequency models has predominantly concentrated on curves, not on probability functions. Although, in principle, both may be transformed into each other, there is an important difference between both approaches: as opposed to curves, the sum of the theoretical (relative) frequencies must be 1, in case of probabilities. Furthermore, the calculation of particular characteristics, such as entropy, repeat rate, etc. is possible only for a system of probabilities, not for curves.
4. The adequacy of a given theoretical model has been tested in different ways: Partly, tests for the goodness of curve approximations (usually the so-called determination R^2) have been applied; partly, however, researchers simply presented tables and/or graphical illustrations, with simple juxtapositions of observed and theoretical values.

In order to guarantee a methodologically consistent procedure, a number of decisions have thus been made with regard to present study:

ad 1: The analysis has been confined to grapheme analyses, only; in how far the conclusions to be drawn are relevant for phoneme studies, as well, will have to be the topic of a separate study. Since by tradition, not all Russian texts use the letter 'ë' as a separate letter in its own right (i.e., identifying 'ë' and 'e'), some texts are composed of 32, others of 33 graphemes; for the present study, therefore, all calculations are based on an inventory size of $n = 32$.

ad 2: Due attention has been paid to the factor of data homogeneity by systematically comparing results obtained on the basis of texts, parts of texts, text cumulations, and text mixtures.

ad 3: With regard to the theoretical model in question, not curves, but probability functions have been applied. In doing so, all relevant models thus far discussed have been tested for their adequacy; therefore, in some cases models containing curve approximations have also been taken into consideration and therefore been transformed into probability models.

ad 4: The goodness of fit has been consequently tested by statistical methods. However, the chi square goodness-of-fit test, which uses to be applied in comparable studies, increases linearly with an increase of sample size. As a result, one is more likely concerned with significant deviations, due to large sample size only (since we are concerned with great sample sizes, in our case). Therefore, it is reasonable to relativize the chi square value by dividing it through the sample size (N) and use the discrepancy coefficient $C = \chi^2/N$, instead; by convention, a value of $C < 0.02$, is interpreted to indicate a good, with $C < 0.01$ a very good fit.

Under these conditions, and meeting the necessary requirements, it will be possible now to test those models discussed in previous research, for the adequacy for Russian grapheme frequencies. Since grapheme systems have a limited number of different classes, however, it seems reasonable, to truncate those distributions the support of which is not $1...n$ (but $1...\infty$) on the right side.

2.1. Zeta Distribution

An early and frequently discussed model is based on consideration by G. K. Zipf. Based on the assumption that the product of the rank (r) of a grapheme and its frequency (f_r) is a constant (c), the resulting equation takes the shape $f_r \times r = c$, which can be represented as

$$(1) \qquad f_r = \frac{c}{r}, \quad r = 1,2,3,...$$

with regard to the theoretical calculation of the frequency. Since formula (1) does not represent a distribution model, however — the theoretical relative frequencies do not sum up to 1, because the harmonic series does not converge, and c is no normalizing constant), it has been enriched by a further parameter (a). The resulting distribution model usually is called Zipf distribution or zeta distribution[4] [Wimmer/Altmann 1999: 664f.]:

$$(2) \qquad P_r = \frac{c}{r^a}, \quad r = 1,2,3,..., \quad a > 1, \quad c^{-1} = \sum_{j=1}^{\infty} \frac{1}{j^a}$$

Truncating this distribution on the right side, one obtains the right-truncated zeta distribution:

$$(3) \qquad P_r = \frac{x^{-a}}{F(R)}, \quad r = 1,2,3,...,R, \quad a \in \square, \quad R \in \square, \quad F(R) = \sum_{i=1}^{R} i^{-a}$$

2.2. Zipf-Mandelbrot Distribution

The generalization of Zipf's original ideas by Mandelbrot results in a more flexible formula with an additional parameter. Based on the initial equation $f_r \times r = c$, this expanded form may be represented as $f_r \times (b + r)^a = c$, which, with regard to the calculation of the theoretical frequencies, leads to

$$(4) \qquad f_r = \frac{k}{(b+r)^a}, \quad r = 1,2,3,...$$

The distribution model resulting from (5) usually is called Zipf-Mandelbrot distribution [Wimmer/Altmann 1999: 666]:

$$(5) \qquad P_r = \frac{c}{(b+r)^a}, \quad r = 1,2,3,..., \quad a > 1, \quad b > -1, \quad c^{-1} = \sum_{j=1}^{\infty} \frac{1}{(b+j)^a}$$

As can be seen, the support of (5) is infinite; therefore truncating it on the right side, one obtains:

$$(6) \qquad P_r = \frac{c}{(b+r)^a}, \quad r = 1,2,3,...,n, \quad a \in \Re, \quad b > -1, \quad c^{-1} = \sum_{j=1}^{n} \frac{1}{(b+j)^a}$$

2.3. Geometric distribution

Another model, which as repeatedly been discussed in context of grapheme frequencies, is based on the so-called geometric series, which consists of the members

$$aq^0, aq^1, aq^2, aq^3, ..., aq^{r-1}, ...$$

In analogy to Zipf's considerations (see above) one can deviate the function

$$(7) \qquad f_r = a \cdot q^r, \quad r = 0,1,2,...$$

from it, which, in the context of grapheme frequencies, has been discussed by Sigurd [1968] or Martindale et al. [1996]. Since in case of ranked frequencies the first rank uses to be denominated as „1" (and not as „0"), and since the function consequently starts with $r = 1$, usually either its 1-displace form

$$(8) \qquad f_r = a \cdot q^{r-1} \qquad r = 1,2,3,...,$$

has been applied, which results in the distribution

$$(9) \qquad P_r = pq^{r-1} \qquad r = 1,2,3,..., \ 0 < q < 1, p = 1-q$$

or the 1-displaced, right-truncated form:

$$(10) \qquad P_r = \frac{(1-q)q^{r-1}}{1-q^n}, \qquad r = 1,2,...,n, \ 0 < q < 1$$

2.4. Good Distribution

Three of the above-mentioned models — namely, (2), (8) und (9) — may be interpreted to be special cases of another distribution which has also been discussed in the context of grapheme frequencies: the so-called Good distribution. Since Good has developed various distribution models, this one has been termed Good1 distribution, in the relevant literature (cf. [Wimmer / Altmann 1999: 219f.]). The Good1 distribution has been brought up by Martindale et al. [1996], who discussed it by reference to the following equation:

$$(11) \qquad P_r = \frac{a}{r^b} \cdot c^r, \quad r = 1,2,...,n$$

In formula (11), a is a normalizing constant which is responsible for the sum of the relative frequencies to sum up to 1:

$$a^{-1} = \sum_{j=1}^{n} \frac{c^j}{j^b}$$

As was pointed out before, there exists a super-ordinate model which explains the relation of distribution (10) to the distributions discussed above. It would lead too far, however, to discuss these interrelations in detail, here (cf. [Grzybek/Kelih/Altmann 2004]). Let it therefore suffice to say that this super-ordinate model is the so-called Lerch distribution (cf. [Zörnig/Altmann 1995]) of which the other models turn out to be special cases. However, thus far this general model has never been applied to grapheme studies.

2. 5. Whitworth Distribution

Another distribution has been discussed by Martindale et al. [1996], though only in form of a curve approximation. In consequently transforming it into a probability density function, Grzybek/Kelih/Altmann [2004] have integrated it into a broader theoretical framework, which need not be presented, here.

It may be sufficient to say that usually, when dealing with distributions, one assumes that the probability of a particular class x or the rank r, develops proportionally to the class below, i.e., to $x - 1$ or $r - 1$, (cf. [Altmann / Köhler 1996]). Based on this general assumption, one obtains the difference equation

$$(12) \qquad P_x = g(x)P_{x-1}$$

the concrete solution of which depends on the particular function $g(x)$. Equation (12) — which, in a way, shows the „top-down" perspective — is rather apt to describe entities with large inventories. If, however, one is concerned with a rather small inventory (i.e., with only a few number of classes), all frequencies have to be balanced in a particular manner, in order to arrive at particular „prescribed" characteristics such as, e.g., entropy or repeat rate). This balance can often be achieved by help of so-called partial sum distributions, which are obtained as follows: if $\{P^*_j\}$ is a given probability function (which

may be termed „base" distribution, for the sake of convention), then one obtains a new distribution:

$$(13) \qquad P_x = C \sum_{j \geq x+k} f(P_j^*)$$

Wimmer / Altmann (2000) have mathematically analyzed various schemes and subsequently discussed with regard to linguistic purposes Wimmer / Altmann [2001]. As far as grapheme analyses are concerned, of the options shown there, only one case has been applied (though without reference to the schemes developed by these two authors), thus far: the partial sum of the discrete rectangular distribution (cf. [Good 1969; Gusein-Zade 1988; Martindale et al. 1996]). The corresponding combinatorial scheme — which has been called 'broken stick distribution', 'distribution of ordered random intervals', or as 'MacArthur distribution' — has been used as early as in the very beginning of the 20th century, however, by Whitworth [1901: 207f.], and shall therefore be called Whitworth distribution, here.

Defining $P_j^* = \frac{1}{n}, \quad j = 1,2,...,n$ (i.e., the discrete rectangular distribution) and applying it to the scheme

$$(18) \qquad P_x = C \sum_{j \geq x} \frac{P_j^*}{j}, \quad x = 1,2,3,...$$

we easily obtain

$$(19) \qquad P_x = \frac{1}{n} \sum_{j=x}^{n} \frac{1}{j}, \quad x = 1,2,3,...,n$$

where $C = 1$. Explicitly writing P_x and adding up, the sum equals 1. It is quite obvious that this distribution has a great advantage: since it has only one parameter (n) which directly results from the inventory size, it can easily be interpreted.

2.6. Negative hypergeometric Distribution

A last distribution model which has to be discussed in the given context, is the so-called negative hypergeometric distribution, which is also known by the name of beta binomial distribution [Wimmer / Altmann 1999: 465ff.]. This model has repeatedly been used for rank frequencies of different kinds: Thus, Köhler / Martináková-Rendeková [1998] have shown that this distribution is an adequate model for ranked frequencies of pitch, intensity, and duration values of a Chopin Étude; and Wimmer / Altmann [2001] and Wimmer / Wimmerová [Ms.], respectively, have successfully modeled rank frequencies of the occurrence of tones in musical works by Bach, Beethoven, Liszt und Chopin.

Thus far, the negative hypergeometric distribution has only rarely been used modeling for linguistic phenomena. Ziegler [2001] has successfully modeled word class frequencies in Portuguese journalistic texts with it. With regard to grapheme frequencies, it

has been used only in a study of A. S. Puškin's *Царь Салтан* [Grzybek 2001]. Since this issue has been further pursued in detail, however, the present study be as well be seen as a more broadly based test of the results obtained on the basis of one text only.

The negative hypergeometric mass function may theoretically derived in various manners, a question which need not be dealt with here. May it suffice to say that its ordinary form

$$(20) \qquad P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}}, \; x = 0,1,2,\ldots,n,\; K > M > 0;\; n \in \{1,2,\ldots\}$$

has to be displaced one step for ranging purposes, what results in its 1-displaced variant

$$(21) \qquad P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}} \qquad x = 1,2,\ldots,n+1;\quad K > M > 0;\; n \in \{1,2,\ldots\}$$

If the negative hypergeometric distribution is truncated at zero, one obtains the positive negative hypergeometric mass function:

$$(22) \qquad P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n} - \binom{K-M+n-1}{n}}, \quad x = 1,2,\ldots,n$$

Interestingly enough, there is a relation between the negative hypergeometric and the Whitworth distribution: In case $K = 2$ and $M = 1$ in (22), one obtains the discrete rectangular distribution, and if one then forms the partial sums as described above, the Whitworth distribution turns out to be a special case of the partially summed negative hypergeometric distribution.

3. *Empirical tests of the models*

3.1. Text and Data Base

As was mentioned above, due attention shall be paid to the important factor of data homogeneity in the present analysis of Russian graphemes. Although this factor is not likely to play a crucial role, in the case of graphemes, a systematic control of this factor seems to be in place, and be it for being "on the safe side" only.

Therefore, the following data material has been used:

a. A first group of texts is represented by complete texts; in order not to follow some a priori definition of 'text', complete chapters of novels have as well been considered to be 'texts' as complete novels. The majority of these texts are literary (prosaic, poetic, and dramatic) texts, for the sake of comparison, technical texts have been included as well.
b. A second group of texts consists of text segments, cumulations, mixtures. Text segments are arbitrarily selected passages of texts, for example particular lines or verses. Text cumulations are successively added chapters of a complete text, which, in the last step, are identical with the complete text. Text mixtures are combinations of arbitrarily texts (or text segments).
c. All texts taken together represent a complete corpus (which, in our case, sums up to ca. 3.3 million graphemes).

Table 1 represents an overview of all analyzed texts and text cumulations. Subsequent to the text number, information about the author or the source of the text can be found, followed by the text's title, its abbreviation, and, finally, its size (in the number of graphemes).

Tab. 1: Text and data basis: complete texts and text cumulations

No.	Author	Text	Chapter	Abbr.	N
1	A.S. Puškin	Evgenij Onegin	1	ASP-EO 1	15830
2			2	ASP-EO 2	11544
3			3	ASP-EO 3	13597
4			4	ASP-EO 4	12475
5			5	ASP-EO 5	12018
6			6	ASP-EO 6	12742
7			7	ASP-EO 7	15180
8			8	ASP-EO 8	15864
9			1-2	ASP-EO 1-2	27374
10			1-3	ASP-EO 1-3	40971
11			1-4	ASP-EO 1-4	53446
12			1-5	ASP-EO 1-5	65464
13			1-6	ASP-EO 1-6	78206
14			1-7	ASP-EO 1-7	93386
15			complete text	ASP-EO 1-8	109250
16	L. N. Tolstoj	*Anna Karenina*	complete text	LNT-AK	1336483
17		*Otročestvo*	complete text	LNT-O	113954
18	F. M. Dostojevskij	*Prestuplenie i nakazanie*	complete text	FMD-PN	837885
19		*Zapiski iz podpol'ja*	complete text	FMD-ZAP	188249
20	A. P. Čechov	*Čajka*	complete text*	APČ-Č	145735
21		*Djadja Vanja*	complete text*	APČ-DV	60871
22	M. Gor'kij	*Mat'*	complete text*	MG-MA	433177
23		*Na dne*	complete text	MG-ND	76039
24	http://www.rusmet.ru/	*Ural'skij rynok metallov*	technischer Text	UR	8061
25	http://www.phyton.ru/	*Instrumental'nye sredstva* [...]	technischer Text	IN	18711

* The dramatic texts signed by an asterisk contain all stage directions, speakers, etc.

Table 2 contains the corresponding data for the text mixtures, text segments, and the complete corpus[5].

Tab. 2: Text and data basis: text mixtures, segments, and corpus

No.	Author	Text	Chapter	Abbr.	N
26	A. S. Puškin	Evgenij Onegin	ch. 1& 8	ASP-EO1+8	31694
27	L. N. Tolstoj	*Anna Karenina*	pt. 8 (ch. 18) & pt. 1 (ch. 1)	LNT-AK8+1	7720
28	F. M. Dostojevskij	*Prestuplenie i nakazanie*	pt. 1 (ch. 1) & pt. 6 (ch. 8)	FMD-PN1+6	29498
29	A. S. Puškin & L. N. Tolstoj	*Evgenij Onegin & Anna Karenina*	complete texts	ASP+LNT	1445733
30	A. S. Puškin & F. M. Dostojevskij	*Evgenij Onegin & Prestuplenie i nakazanie*	complete texts	ASP+FMD	947135
31	A. S. Puškin & text 24	*Evgenij Onegin & Text 24*	complete texts	ASP+UR	117311
32	L. N. Tolstoj & text 24	*Anna Karenina & Text 24*	complete texts	LNT+UR	1344544
33	F. M. Dostojevskij & text 25	*Prestuplenie i nakazanie & Text 25*	complete texts	FMD+IN	856596
34	M. Gor'kij & text 25	*Na dne & Text 25*	complete texts	MG+IN	95312
35	Puškin, A. S.	*Evgenij Onegin*	ch. 5, verse 1-5 per ch.	ASP1-5	4323
36	F. M. Dostojevskij	*Prestuplenie i nakazanie*	epilogue, each alternate line	FMD-2	14464
37	L. N. Tolstoj	*Anna Karenina*	pt. 4 (ch. 1—5), every 4th line	LNT-4	7141
38	Complete corpus			CC	3328454

3.2. Results

Let us now take a look at the results for the six models discussed above. The basic

Table 3 presents the absolute frequenc
1 to 32. Whereas parameter $n = 32$ corre
b are the result of theoretical estimations
obtains $a = 1.5281$ and $b = 5.2679$. Fillin
theoretical values $NP(i)$:

i	f(i)	NP(i)
1	982048	24160,
2	763584	19269,
3	701891	15823,
4	593949	13290,
5	563851	11363,
6	532783	9859,
7	456610	8658,
8	423657	7680,
9	403285	6873,
10	353818	6197,
11	295548	5625,
12	272216	5134,
13	262459	4711,
14	248196	4343,
15	222221	4019,
16	195629	3734,
	$a = 1.5281$	
	$b = 5.2679$	
	$n = 32$	

As can be seen from the values in table 3 and their graphic representation in fig. 1, the Zipf-Mandelbrot is no good model, in this case. This fact is corroborated by the poor value of $C = 0.0492$. The very same tendency holds true for all other cases, as well; and, since the Zipf-Mandelbrot distribution is a generalization of the zeta distribution, the latter also turns out to be no adequate model for Russian grapheme frequencies. Table 4a/b represents the results of fitting the zeta and the Zipf-Mandelbrot distributions to all data sets.

It can clearly be seen that both distribution which have repeatedly been applied for grapheme frequencies, are not really adequate for modeling rank frequencies of Russian graphemes. In case of the zeta distribution, the values of the discrepancy coefficient are in the interval $0.1664 \geq C \geq 0.0995$, for the complete corpus it is $C = 0.1177$ — not a single sample arrives at a value of $C < 0.02$. The Zipf-Mandelbrot distribution, too, which has one more parameter (a, b, n) as compared to the zeta distribution, does not seem to be an appropriate model: Only three of the samples arrive at a discrepancy coefficient of $C < 0.02$. Interestingly enough, however, the value of the corpus is relatively satisfying, as compared to the results for the individual samples, which asks for further investigation. Fig. 2 illustrates the results, showing the discrepancy coefficients for all samples.

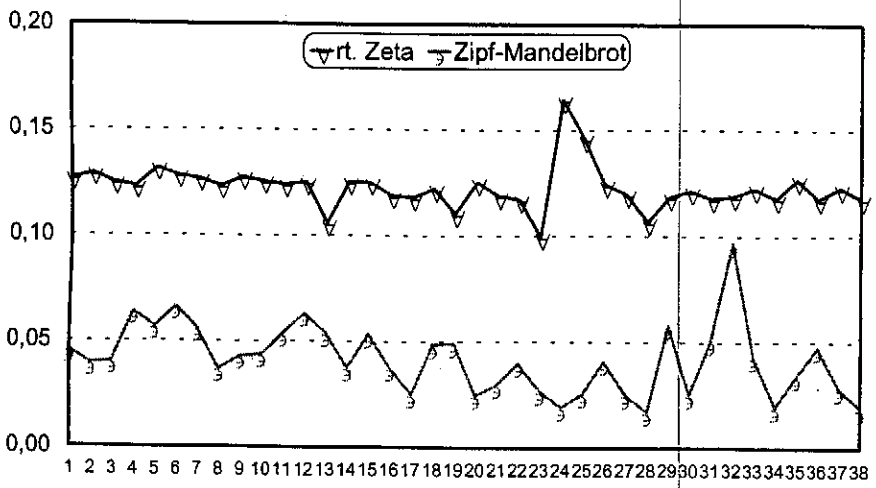


Fig. 2: Discrepancy coefficients of the zeta and Zipf-Mandelbrot distributions

Both models therefore can be ruled out from further considerations which shall concentrate on the right truncated geometric and the right truncated Good distribution in the next step. Table 5/a/b shows the results in detail.

Tab. 4a/b: Right truncated zeta and Zipf-Mandelbrot distribution

		right truncated zeta, R=32			Zipf-Mandelbrot (a,b) n=32			
No.	Abbr.	a	$\chi^2_{FG:29}$	C	a	b	$\chi^2_{FG=28}$	C
1	ASP-EO 1	0,6659	2001,81	0,1265	2,2784	12,8065	721,25	0,0456
2	ASP-EO 2	0,6828	1487,25	0,1288	2,7371	16,5558	456,80	0,0396
3	ASP-EO 3	0,6725	1693,67	0,1246	2,7403	17,1914	551,15	0,0405
4	ASP-EO 4	0,6714	1530,24	0,1227	1,4136	5,0791	794,37	0,0637
5	ASP-EO 5	0,6703	1580,15	0,1315	1,7478	7,8099	681,15	0,0567
6	ASP-EO 6	0,6728	1632,08	0,1281	1,3574	4,4606	842,83	0,0661
7	ASP-EO 7	0,6722	1920,19	0,1265	1,6214	6,6847	847,98	0,0559
8	ASP-EO 8	0,6886	1951,23	0,1230	3,0305	19,4182	587,00	0,0370
9	ASP-EO 1-2	0,6726	3481,37	0,1272	2,4360	14,0741	1177,47	0,0430
10	ASP-EO 1-3	0,6725	5128,28	0,1252	2,2610	12,5103	1800,41	0,0439
11	ASP-EO 1-4	0,6721	6622,94	0,1239	1,6966	7,4194	2859,07	0,0535
12	ASP-EO 1-5	0,6710	8179,52	0,1249	1,3976	4,8765	4121,98	0,0630
13	ASP-EO 1-6	0,7172	8257,16	0,1056	1,7176	7,5936	4196,80	0,0537
14	ASP-EO 1-7	0,6714	11690,57	0,1252	3,0800	20,3454	3515,08	0,0376
15	ASP-EO 1-8	0,6737	13646,19	0,1249	1,7081	7,4547	5825,38	0,0533
16	LNT-AK	0,7238	158702,58	0,1187	2,1424	9,9287	50451,88	0,0377
17	LNT-OT	0,6991	13438,62	0,1179	6,5497	52,2365	2852,85	0,0250
18	FMD-PR	0,7113	102121,20	0,1219	1,7772	7,1180	40547,22	0,0484
19	FMD-ZA	0,7119	20745,41	0,1102	1,5282	5,2679	9265,88	0,0492
20	APČ-ČA.	0,7073	18228,33	0,1251	12,0000	118,9459	3586,48	0,0246
21	APČ-DJ.	0,7129	7226,2866	0,1187	2,6542	14,8642	1813,34	0,0298
22	MG-MA.	0,7046	50716,0098	0,1171	2,3162	11,9652	17278,67	0,0399
23	MG-NA	0,6982	7563,71	0,0995	7,6666	65,0316	2028,44	0,0267
24	UR	0,7330	1325,26	0,1644	12,0000	102,3936	155,05	0,0192
25	IN	0,7098	2731,32	0,1460	5,6544	40,2040	478,15	0,0256
26	ASP-EO1+8	0,6766	3938,82	0,1243	2,5939	15,5644	1298,49	0,0410
27	LNT-AK8+1	0,7232	924,12	0,1197	5,1200	36,4961	194,05	0,0251
28	FMD-PR1+6	0,6993	3147,01	0,1067	10,8361	94,8234	504,90	0,0171
29	ASP+LND	0,7200	171135,22	0,1184	12,0000	144,0974	83390,77	0,0577
30	ASP+FMD	0,7068	114782,89	0,1212	5,7090	43,1875	24000,24	0,0253
31	ASP+UR	0,6370	13793,61	0,1176	1,8231	8,3289	5941,93	0,0507
32	LNT+UR	0,7239	159407,53	0,1186	12,0000	167,6190	129723,30	0,0965
33	FMD+IN	0,7108	104658,16	0,1222	2,0404	9,2915	36928,53	0,0431
34	MG+IN	0,6868	11190,42	0,1175	12,0000	117,0180	1833,61	0,0192
35	ASP1-5	0,6901	546,93	0,1265	4,4701	33,2901	147,36	0,0341
36	FMD-2	0,7336	1698,62	0,1174	1,7018	6,1837	691,26	0,0478
37	LNT-4	0,7265	876,40	0,1227	2,9291	16,4118	202,65	0,0284
38	CC	0,7133	391831,09	0,1177	12,0000	105,0550	13874,96	0,0190

Tab. 5a/b: Right truncated geometric right truncated Good-1 distributions

		right truncated geometric, $R = 32$			right truncated Good-1 (a, p)			
No.	Text	q	$\chi^2_{FG=29}$	C	a	p	$\chi^2_{FG=29}$	C
1	ASP-EO 1	0,9086	443,91	0,0280	0,00000014	0,89976	471,65	0,0298
2	ASP-EO 2	0,9064	313,89	0,0272	0,00000030	0,89811	328,77	0,0285
3	ASP-EO 3	0,9083	399,39	0,0294	0,00000008	0,89966	422,38	0,0311
4	ASP-EO 4	0,9087	420,24	0,0337	0,74387629	0,98000	1716,30	0,1376
5	ASP-EO 5	0,9078	354,32	0,0295	0,00000002	0,89914	373,88	0,0311
6	ASP-EO 6	0,9072	337,20	0,0265	0,00000001	0,89867	356,44	0,0280
7	ASP-EO 7	0,9076	394,38	0,0260	0,74717027	0,98000	2115,18	0,1393
8	ASP-EO 8	0,9064	463,37	0,0292	0,00003339	0,89819	483,07	0,0305
9	ASP-EO 1-2	0,9077	752,14	0,0275	0,00000025	0,89911	795,07	0,0290
10	ASP-EO 1-3	0,9080	1122,48	0,0274	0,00000000	0,89936	1189,19	0,0290
11	ASP-EO 1-4	0,9082	1515,85	0,0284	0,00000018	0,89950	1606,65	0,0301
12	ASP-EO 1-5	0,9083	1878,70	0,0287	0,00000001	0,89956	1988,54	0,0304
13	ASP-EO 1-6	0,9081	2195,71	0,0281	0,00000031	0,89943	2324,36	0,0297
14	ASP-EO 1-7	0,9080	2563,15	0,0274	0,00000016	0,89931	2718,88	0,0291
15	ASP-EO 1-8	0,9078	3015,62	0,0276	0,00000007	0,89918	3188,81	0,0292
16	LNT-AK	0,9002	28735,24	0,0215	0,80841128	0,98000	181444,61	0,1358
17	LNT-OT	0,9038	2734,34	0,0240	0,80619800	0,98000	16828,51	0,1477
18	FMD-PR	0,9003	17699,49	0,0211	0,77618992	0,98000	108400,81	0,1294
19	FMD-ZA	0,9025	4464,71	0,0237	0,00000951	0,89501	4589,25	0,0244
20	APČ-ČA.	0,9034	3271,84	0,0225	0,77321643	0,98000	19431,64	0,1333
21	APČ-DJ.	0,9027	1200,12	0,0197	0,00000001	0,89528	1240,04	0,0204
22	MG-MA.	0,9028	10826,33	0,0250	0,78065173	0,98000	57155,47	0,1319
23	MG-NA	0,9063	2039,73	0,0268	0,00000000	0,89805	2134,88	0,0281
24	UR	0,8940	134,70	0,0167	0,80235312	0,98000	1309,22	0,1624
25	IN	0,8987	358,76	0,0192	0,78144672	0,98000	2787,69	0,1490
26	ASP-EO1+8	0,9076	904,82	0,0285	0,00000007	0,89137	1388,15	0,0438
27	LNT-AK8+1	0,8997	173,62	0,0225	0,78516143	0,98000	972,70	0,1260
28	FMD-PR1+6	0,9043	561,27	0,0190	0,75336999	0,98000	3298,06	0,1118
29	ASP+LNT	0,9008	30774,89	0,0213	0,78285514	0,98000	181446,72	0,1255
30	LNT+FMD	0,9013	20958,83	0,0221	0,78699029	0,98000	128874,45	0,1361
31	ASP+UR	0,9070	3071,67	0,0262	0,00000002	0,89853	3240,76	0,0276
32	LNT+UR	0,9002	28774,37	0,0214	0,78783364	0,98000	169893,70	0,1264
33	FMD+IN	0,9013	19026,55	0,0222	0,77609868	0,98000	111138,41	0,1297
34	MG+IN	0,9061	1923,35	0,0202	0,77619024	0,98000	13128,46	0,1378
35	ASP1-5	0,9061	132,49	0,0306	0,75891235	0,98000	596,99	0,1381
36	FMD-2	0,8987	357,04	0,0247	0,00000015	0,89200	359,32	0,0248
37	LNT-4	0,8990	123,35	0,0173	0,00000070	0,89217	124,93	0,0175
38	CC	0,9016	69203,68	0,0208	0,777321050	0,98000	41751,32	0,1254

As the results presented in table 5a/b show, neither the geometric nor the Good distribution yield satisfying results. For the geometric distribution, the values of the discrepancy coefficient are in the interval $0.337 \geq C \geq 0.0167$ for the individual samples, of which only five have a value of $C < 0.02$; as to the corpus, the value of $C = 0.208$ is better than the one for most of the individual sample, but still fails the level of significance. The results are even worse for the Good distribution; in only one case, C underscores the level of significance. Fig. 3 illustrates the results, showing the discrepancy coefficients for all samples.

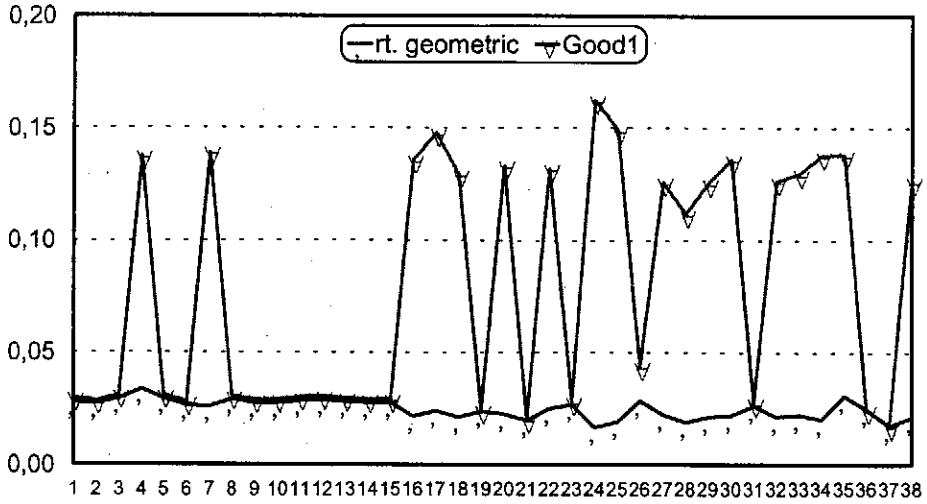


Fig. 3: Discrepancy coefficients of the geometric and Good distributions

As can be seen from fig. 3, the values of the Good distribution are far from being stabile; still, there is no plausible explanation (sample size, data homogeneity, authorship, text type, etc.) for this tendency. As compared to this, the geometric distribution is rather stabile, but constantly above the level of significance. We can thus rule out those four (of our six) models, which have predominantly been applied in previous research. This fact makes it even more important to look for new ways, and to test the remaining two models, in order to find out, if either the Whitworth or the negative hypergeometric distribution yields better results.

As was mentioned above, the Whitworth distribution is particularly attractive due to the fact that it has one parameter (n) which may easily be interpreted. Therefore, it is of utmost importance to note that this distribution yields satisfying results, as far as ranked frequencies of Russian graphemes are concerned. Table 8 represents the results in detail:

Tab. 6: Whitworth distribution

Whitworth, R = 32							
No.	**Text**	$\chi^2_{FG=30}$	**C**	**No.**	**Text**	$\chi^2_{FG=30}$	**C**
1	ASP-EO 1	275,79	0,0174	**20**	APČ-ČA.	2075,78	0,0142
2	ASP-EO 2	196,62	0,0170	**21**	APČ-DJ.	631,01	0,0104
3	ASP-EO 3	252,01	0,0185	**22**	MG-MA.	3880,32	0,0090
4	ASP-EO 4	245,82	0,0197	**23**	MG-NA	1230,06	0,0162
5	ASP-EO 5	227,89	0,0190	**24**	UR	170,95	0,0212
6	ASP-EO 6	201,05	0,0158	**25**	IN	351,32	0,0188
7	ASP-EO 7	246,78	0,0163	**26**	ASP-EO1+8	526,22	0,0166
8	ASP-EO 8	255,72	0,0161	**27**	LNT-AK8+1	57,30	0,0074
9	ASP-EO 1-2	465,81	0,0170	**28**	FMD-PR1+6	236,22	0,0080
10	ASP-EO 1-3	692,18	0,0169	**29**	ASP+LNT	11508,38	0,0080
11	ASP-EO 1-4	908,01	0,0170	**30**	ASP+FMD	8461,09	0,0089
12	ASP-EO 1-5	1142,42	0,0175	**31**	ASP+UR	1798,70	0,0153
13	ASP-EO 1-6	1316,45	0,0168	**32**	LNT+UR	10539,38	0,0078
14	ASP-EO 1-7	1536,60	0,0165	**33**	FMD+IN	7135,16	0,0083
15	ASP-EO 1-8	1784,02	0,0163	**34**	MG+IN	1128,00	0,0118
16	LNT-AK	10464,05	0,0078	**35**	ASP1-5	78,00	0,0180
17	LNT-OT	1094,06	0,0096	**36**	FMD-2	113,15	0,0078
18	FMD-PR	6831,59	0,0082	**37**	LNT-4	54,30	0,0076
19	FMD-ZA	1243,79	0,0066	**38**	CC	22763,51	0,0068

The results presented in table 6 clearly prove the Whitworth distribution to be an adequate model for ranked grapheme frequencies in Russian: for the indidual samples, the discrepancy coefficient is in the interval $0.212 \geq C \geq 0.0066$; for 23 of the 37 samples, the discrepancy coefficient is $C < 0.02$, in 13 cases even $C < 0.01$ — only one of the texts (the technical text #28) slightly fails the defined level of significance.

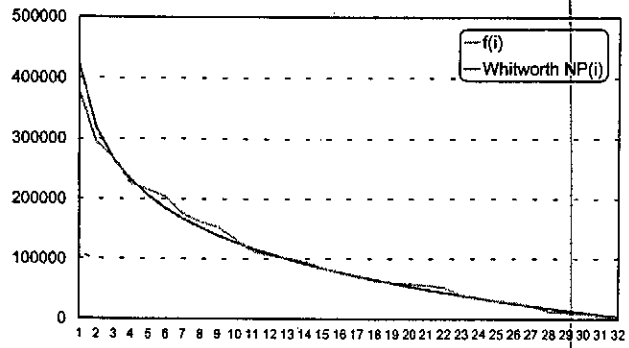


Fig. 4a: Fitting the Whitworth distribution (complete corpus)

Fig. 4a graphically presents the good fitting result for the whole corpus ($C = 0.0068$); the overall stability of the discrepancy coefficient C for all 38 data sets, is illustrated in fig. 4b.

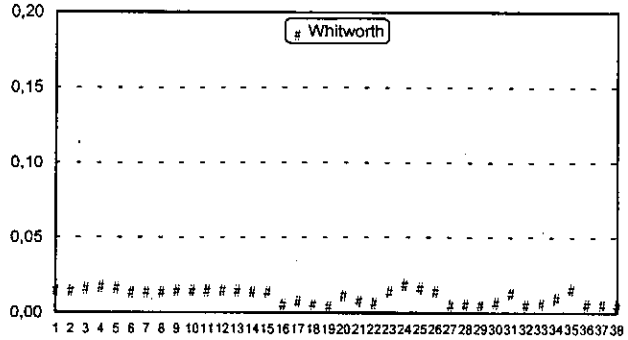


Fig. 4b: Constancy of the discrepancy coefficient *C* for fitting the Whitworth distribution (complete corpus)

Since the results for the negative hypergeometric distribution are even better, they shall be presented here in detail, arriving at an end of our study. Table 6 / fig. 5 present the results for the complete corpus.

Tab. 6: Negative hypergeometric distribution (corpus)

i	**f(i)**	**NP(i)**	**i**	**f(i)**	**NP(i)**
1	377272	392798,11	**17**	69666	74399,29
2	293471	299110,55	**18**	62778	67406,45
3	269118	256811,89	**19**	59872	60752,49
4	225809	228556,95	**20**	58037	54418,89
5	215165	206814,32	**21**	55743	48391,21
6	203776	188840,80	**22**	52166	42658,63
7	175642	173340,29	**23**	38918	37213,55
8	162044	159600,48	**24**	34719	32051,45
9	153632	147188,88	**25**	30119	27170,98
10	133968	135823,67	**26**	26924	22574,13
11	113077	125311,27	**27**	21071	18266,81
12	104465	115513,27	**28**	11913	14259,90
13	100794	106327,52	**29**	9558	10571,06
14	95430	97676,72	**30**	8442	7228,37
15	84749	89501,25	**31**	3845	4277,90
16	75453	81754,42	**32**	779	1803,51
K = 3,1511				c^2 = 13874,96	
M = 0,7948				FG = 28	
n = 31				C = 0,0042	

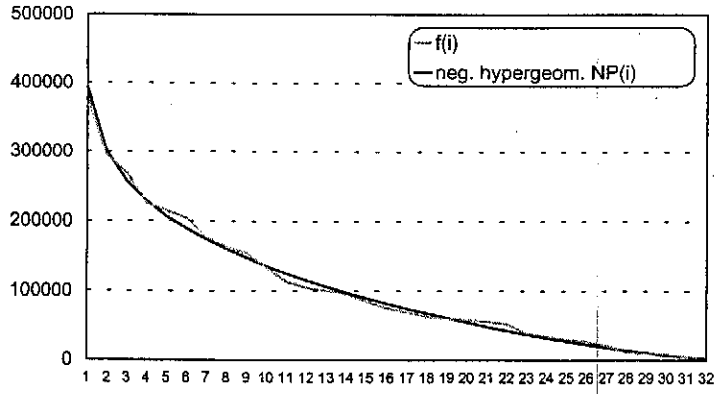


Fig. 5: Fitting the negative hypergeometric distribution (complete corpus)

A comparison with the results for all individual sample confirms the impression that the negative hypergeometric distribution is an excellent model for ranked grapheme frequencies in Russian: in all cases, the discrepancy coefficient is in the interval $0.0169 \geq C \geq 0.0043$, and in not less but 32 of the 37 individual samples the discrepancy coefficient is not only $C < 0.02$, but even $C < 0.01$.

Tab. 7: Negative hypergeometric distribution

Neg. Hypergeometric, n=31					
No.	**Text**	**K**	**M**	$\chi^2_{EG=28}$	**C**
1	ASP-EO 1	3,1904	0,8472	85,94	0,0054
2	ASP-EO 2	3,2120	0,8394	99,87	0,0087
3	ASP-EO 3	3,1751	0,8405	114,06	0,0084
4	ASP-EO 4	3,1388	0,8306	118,41	0,0095
5	ASP-EO 5	3,2388	0,8531	87,83	0,0073
6	ASP-EO 6	3,2061	0,8450	67,31	0,0053
7	ASP-EO 7	3,2001	0,8445	89,34	0,0059
8	ASP-EO 8	3,1666	0,8250	148,69	0,0094
9	ASP-EO 1-2	3,1974	0,8439	178,68	0,0065
10	ASP-EO 1-3	3,1853	0,8422	269,22	0,0066
11	ASP-EO 1-4	3,1742	0,8397	356,24	0,0067
12	ASP-EO 1-5	3,1816	0,8418	445,28	0,0068
13	ASP-EO 1-6	3,1868	0,8429	480,76	0,0061
14	ASP-EO 1-7	3,1894	0,8434	541,67	0,0058
15	ASP-EO 1-8	3,1869	0,8411	679,85	0,0062
16	LNT-AK	3,1412	0,7893	8231,12	0,0062
17	LNT-OT	3,1084	0,8015	594,66	0,0052
18	FMD-PR	3,1567	0,8005	3839,72	0,0046
19	FMD-ZA	3,0454	0,7818	805,17	0,0043
20	APČ-ČA.	3,1644	0,8137	1691,75	0,0116
21	APČ-DJ.	3,1245	0,8050	533,09	0,0088
22	MG-MA.	3,1065	0,7959	2165,86	0,0050
23	MG-NA	3,1563	0,8259	765,06	0,0101
24	UR	3,4201	0,8269	120,06	0,0149
25	IN	3,1483	0,7927	316,62	0,0169
26	ASP-EO1+8	3,1736	0,8356	229,95	0,0073
27	LNT-AK8+1	3,1401	0,7872	38,47	0,0050
28	FMD-PR1+6	2,9490	0,7707	149,17	0,0051
29	ASP+LNT	3,1400	0,7909	8830,29	0,0061
30	ASP+FMD	3,1465	0,8027	4841,49	0,0051
31	ASP+UR	3,2017	0,8410	686,02	0,0058
32	ASP+UR	3,1422	0,7894	8288,54	0,0062
33	FMD+IN	3,1542	0,8004	4079,82	0,0048
34	MG+IN	3,1014	0,8161	580,80	0,0061
35	ASP1-5	3,1854	0,8282	49,02	0,0113
36	FMD-2	3,1524	0,7816	87,49	0,0060
37	LNT-4	3,1651	0,7910	46,20	0,0065
38	CC	3,1441	0,7948	13874,96	0,0042

As can be seen from the results presented in table 7, the discrepancy coefficient C turns out to be convincingly stabile for all samples. Interestingly enough, the parameters, too, display a convincing degree of stability. This holds true not only for parameter n (which is defined by the inventory size minus one and therefore is constantly $n = 31$), also the values for K and M are extremely stabile: $3.42 \geq K \geq 2.95$ und $0.85 \geq M \geq 0.77$. Figs. 6a/b illustrate the constancy of the results.

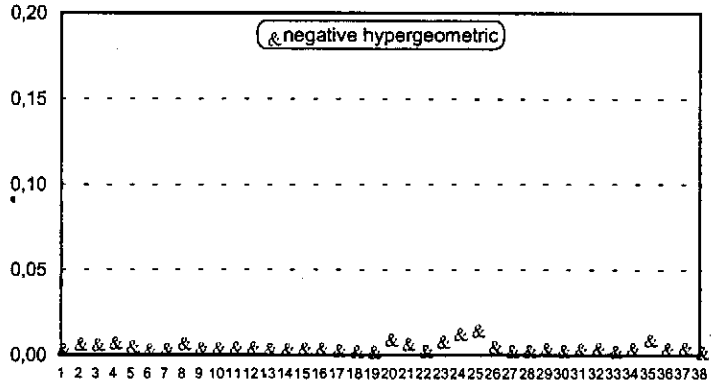


Fig. 6a: Constancy of the discrepancy coefficient

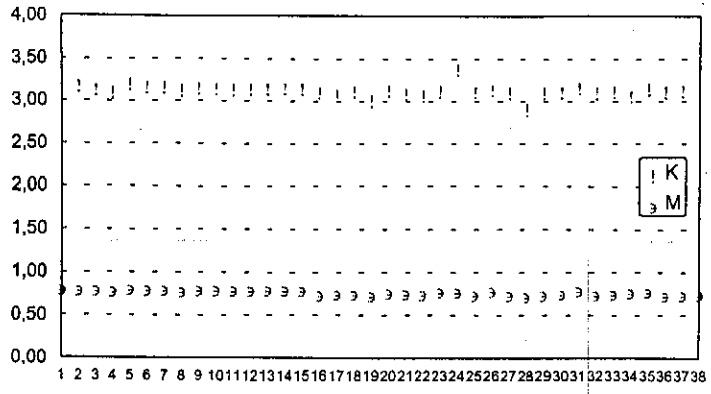


Fig. 6b: Constancy of the parameters *K* und

5. Summary, Conclusion, Perspectives

The results of the present study allow for a number of conclusions, which pave the way for further research:

1. The Russian grapheme system seems to be an orderly organized system, as far as the frequency of its elements are concerned.
2. The fact of the law-like organization of the Russian grapheme system allows for the extended hypothesis that the elements of the graphemes may be systematically organized, as well; with regard to this question, both theoretical and empirical work is needed.
3. The question of data homogeneity obviously plays only a minor, if any role on the level of grapheme analyses. Although in some cases, the results obtained were better for the complete corpus than for individual cases, 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.
4. It is an important finding that four distributions, which have been assumed to be adequate models in previous research (zeta, Zipf-Mandelbrot, geometric, Good) are not acceptable for Russian grapheme frequencies; most likely, a number of assumptions about other languages will have to be modified.
5. In case of Russian, a relative easy and plausible model, the Whitworth distribution, yields satisfying results, although this model has rarely been used in previous research. Future studies will have to show in how far grapheme frequencies of other languages, too, can be covered by this model; preliminary data from other Slavic languages show that the Whitworth distribution does not seem to be of some general validity (cf. [Grzybek/Kelih 2003b] for Slovene, and [Benko / Grzybek / Kelih / Kusendova / Nemcová 2004] for Slovak).

6. The negative hypergeometric distribution turns out to be an excellent model. 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.
7. The extension of the studies to other Slavic languages may provide insights not only as to their graphemic structure(s), in general, but also as to historical-diachronic aspects of this question.
8. In further pursuing this line of research, it will be of utmost importance to search for cross-references with the corresponding phonemic systems. 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.
9. In addition to the extension of this research to other (Slavic) languages, a more detailed theoretical treatment of the discussed models will be necessary; notwithstanding the fact, that some transitions between the models discussed here and overall generalizations have already been described by Grzybek/Kelih/Altmann [2004], it will be necessary to elaborate specific characteristics such as, e.g., the theoretical entropies and repeat rates of these models in order to arrive at solid conclusions.

Summarizing, one can say that the study of the Russian graphemic system (as of other graphemic systems, too, of course) definitely goes beyond the mere counting of letters, and that it is by no way a matter of the past — rather there are still many perspectives for future theoretical and empirical research, which in part, have been outlined some decades ago.

NOTES

[1] Unfortunately, the text by Nikolaeva [1961a] was not available to the author of these lines; it seems that history repeats itself, though with reversed premises, if one reads Nikolaeva's [1965: 130] words, written some decades ago: "I regard it necessary to remark that some works on this question, which are known from bibliographical resources (...), remained unstudied due to reasons of technical character."

[2] In fact, Volockaja et al. [1964: 14ff.] arrive at six differences, choosing a different representation for the letters 'д' and 'г', namely the italic forms *д* and *г*.

[3] Interestingly enough, almost the same elements pointed out by Volockaja et at. (1964) have been distinguished by W. A. Koch [1971: 74f.] in his attempt to describe the Latin alphabet (in its variant for the English grapheme system). Additionally, however, Koch has given a number of additional "features" which, in fact, are rather considered to be either positional specifications (such as, 'top', 'bottom', 'middle'), or particular rules for handling the basic elements (such as, e.g. 'reduced', 'mirrored', or 'crossed'). — Meanwhile, there are quite a number of attempts to describe alphabetic systems, all of them arriving at different solutions, and it would be worthwhile checking their relevance for the detailed description of Russian letters, too — cf. Althaus [1973: 108], Boudon [1981: 35ff.], Mounin [1970: 135ff.], Watt [1975; 1978; 198]; 1988; 2002].

[4] The name zeta distribution goes back to the fact that in equation (2) $k^1 = \zeta(a)$ is Riemann's zeta function; this distribution has a number of further names, however, such as discrete Pareto distribution, Joos model, Riemann's zeta distribution, Zipf-Estoup distribution, Zipf's Law, etc. (cf. [Wimmer / Altmann 1999: 664f.]).

[5] As opposed to the text by Grzybek/Kelih [2004], the term «corpus» refers to the totality of all complete texts, only (thus not counting text segments, mixtures, and cumulations more than once in data set #38; therefore, the results are slightly different as compared to the mentioned study..

References

Althaus, H. P. (1973): «Graphetik». In: H. P. Althaus; H. Henne; H. E. Wiegand (eds.), *Lexikon der Germanistischen Linguistik*. Tübingen. (105—110).

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

Benko, V. / Grzybek, P. / Kelih, E. / Kusendová, J. / Nemcová, E. (2004): «Rank Frequency Models for Slowak Graphemes». [In prep.]

Boudon, P (1981): *Introduction à une sémiotique des lieux*. Paris.

Catford, J. C. (1965): *A Linguistic Theory of Translation*, London.

Good, I. J. (1969): «Statistics of Language: Introduction». In: Meetham, C. A.; Hudson, R. A. (eds.), *Encyclopaedia of Linguistics, Information, and Control*. Oxford etc. (567—581).

Grzybek, P. (2001): «Kultur — Ökonomie. Zur Häufigkeit text-konstitutiver Elemente». In: Weitlaner, W. (Hg.), *Sprache — Kultur — Ökonomie*. Wien. (485—509). [= Wiener Slawistischer Almanach, Sonderband 54]

Grzybek, P.; Kelih, E. (2003a): «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. (2003b): «Grapheme frequencies in Slovene». In: V. Benko (ed.), *Slovko 2003*. Bratislava. [In print]

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].

Gusein-Zade, S. M. (1988): «O raspredelenii bukv russkogo jazyka po častote vstrečaemosti». In: *Problemy peredači informacii, 24*,; 102—107.

Koch, W. A. (1971): *Taxologie des Englischen*. München.

Köhler, R.; Martináková-Rendeková, Z. (1998): «A systems theoretical approach to language and music». In: Altmann, G.; Koch, W. A. (eds.) (1998): *Systems. New Paradigms for the Human Sciences*. Berlin / New York: de Gruyter. 514—546.

Martindale, C.; Gusein-Zade, S. M.; McKenzie, D.; Borodovsky, M. Yu. (1996): «Comparison of Equations Describing the Ranked Frequency Distributions of Graphemes and Phonemes». In: *Journal of Quantitative Linguistics, 3*,; 106—112.

Mounin, G. (1970): *Introduction à la sémiologie*. Paris.

Nikolaeva, T. M. (1961a): «Klassifikacija russkich grafem». In: *Doklady na konferencii po obrabotke informacii, mašinnomu perevodu, i avtomatičeskomu čteniju, 6*. Moskva.

Nikolaeva, T. M. (1961b): «Pis'mennja reč' i specifika ee izučenija», in: *Voprosy jazykoznanija, 3*; 78—86.

Nikolaeva, T. M. (1965): «Čto takoe grafema?». In: *Filologičeskie nauki, 3*; 130—134.

Nikolaeva, T. M. (1969): «Problemy opisanija edinic plana vyraženija: Sintez čerez analiz». In: *Trudy po znakovym sistemam IV*. Tartu. (483—486).

Sigurd, B. (1968): «Rank-Frequency Distributions for Phonemes». In: *Phonetica, 18*; 1—15.

Volockaja, Z. M.; Mološnaja, T. N. Nikolaeva, T. M. (1964): *Opyt opisanija russkogo jazyka v ego pis'mennoj forme*. Moskva.

Watt, W. C. (1975):«What is the proper characterization of the alphabet? Part I: Desiderata». In: *Visible Language, 9*; 293—327.

Watt, W. C. (1980): «What is the proper characterization of the alphabet? Part II: Composition». In: *Ars Semeiotica, 3*; 3—46.

Watt, W. C. (1981): «What is the proper characterization of the alphabet? Part III: Appearance». In: *Ars Semeiotica, 4*; 269—313.

Watt, W. C. (1988): «What is the proper characterization of the alphabet? Part IV: Union». In: *Semiotica, 70*; 199—241.

Watt, W. C. (1988): «What Is the proper characterization of the Alphabet? V: Transcendence». In: *Semiotica, 138*; 131—178.

Whitworth, W. A. (1901): *Choice and Chance. With One Thousand Exercises*. New York / London: Hafner, 1965.

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

Wimmer, G.; Altmann, G. (2000): «On the Generalization of the STER Distribution Applied to Generalized Hypergeometric Parents». In: *Acta Universitatis Palackianae Olomucensis, Facultas rerum naturalium, Mathematica, 39*; 215—247.

Wimmer, G.; Altmann, G. (2001): «Models of Rank-Frequency Distributions in Language and Music». In: L. Uhlířová; G. Wimmer, G. Altmann, R. Köhler (eds.), *Text as a Linguistic Paradigm: Festschrift in honour of Luděk Hřebíček*. Trier: WVT. (283—294).

Wimmer, Gejza; Wimmerová, Soňa (Ms.): «Ein musikalisches Rangordnungsgesetz».

Ziegler, A. (2001): «Word Class Frequencies in Portuguese Press Texts». In: L. Uhlířová; G. Wimmer, G. Altmann, R. Köhler (eds.), *Text as a Linguistic Paradigm: Festschrift in honour of Luděk Hřebíček*. Trier: WVT. (295—312).

Zörnig, P.; Altmann, G. (1995): «Unified representation of Zipf distributions». In: *Computational Statistics & Data Analysis 19, 461—473*.

ББК 81.2Рус-67-1
Я 41

Издание осуществлено при финансовой поддержке
Российского фонда фундаментальных исследований
(РФФИ)
проект № 04-06-87066

Я 41 Язык. Личность. Текст: Сб. ст. к 70-летию Т. М. Николаевой / Ин-т славяноведения РАН; Отв. ред. В. Н. Топоров. — М.: Языки славянских культур, 2005. — 976 с. — (Studia philologia).

ISSN 1726-135X
ISBN 5-9551-0103-9

Сборник посвящен юбилею члена-корреспондента РАН Т. М. Николаевой. В нем публикуются статьи по теории языкознания, по проблемам грамматики, фонетики и интонологии, по семиотике и мифологии, а также по литературоведению. Многообразие тем отражает широту научных занятий и интересов юбиляра.

ББК 81.2Рус

ISBN 5-9551-0103-9

ИНСТИТУТ СЛАВЯНОВЕДЕНИЯ РАН

ЯЗЫК ЛИЧНОСТЬ ТЕКСТ

СБОРНИК СТАТЕЙ
К 70-ЛЕТИЮ
Т. М. НИКОЛАЕВОЙ

Ответственный редактор
В. Н. Топоров

ЯЗЫКИ СЛАВЯНСКИХ КУЛЬТУР
МОСКВА 2005