Mathematical Notes, vol. 80, no. 5, 2006, pp. 679–691.
Translated from Matematicheskie Zametki, vol. 80, *no.* 5, 2006, *pp.* 718–732.
Original Russian Text Copyright © 2006 *by V. P. Maslov, T. V. Maslova.*

On Zipf's Law and Rank Distributions in Linguistics and Semiotics

V. P. Maslov and T. V. Maslova

Received September 27, 2006

Abstract—A number of formulas of linguistic statistics are refined. The notions of real and virtual cardinality of a sign are introduced. We show that a formula refining Zipf's law for the occurrence frequencies in frequency dictionaries can be extended to arbitrary sign objects, i.e., semiotic systems.

Key words: *linguistic statistics*, *real and virtual cardinality of a sign*, *Zipf's law*, *Bose–Einstein distribution*, *semiotic system*, *frequency dictionary*, *occurrence frequency.*

1. SYNTAGMATIC (ORDERED) WORD SEQUENCES AND WRITERS' MARKERS

Kolmogorov, Dobrushin, Gusein-Zade, and many other mathematicians and linguists were interested in the mathematical aspects of language structure and functioning. Kolmogorov wrote several papers on the statistical laws of prosody. Gusein-Zade, in the spirit of Kolmogorov's and Shannon's ideas, treated alphabetic letter frequencies as a marker of the language [1]. It is natural to ask whether word occurrence frequencies can be viewed as a marker (i.e., identifier) of the language of individual writers.

The text authorship attribution problem, especially for Shakespearian texts (for which a prize of over £1 million has been offered [2]), has attracted the attention of scientists for centuries. The renowned scientist N. A. Morozov (who was also a member of People's Will revolutionary group) and his followers tried to hand-compute the frequencies of auxiliary words; however, Morozov's paper [3] was refuted by the famous mathematician Markov [4]. Thus, Markov himself became interested in this problem.

We doubt that authorship can indeed be attributed on the basis of word frequencies, but the surprising coincidence of the word frequency statistics for various authors with the formulas given below permit one to define a parameter that depends on the text length (i.e., a parameter–text length curve) and is a well-defined characteristic of a given text.

Text statistics laws and Zipf's law were studied by Arapov, Efimova, and Shreider [5], who noticed that, while relatively small texts show a good agreement with Zipf's law, the law is violated for long texts consisting of many relatively independent parts. It is natural to assume that, when generating the text, the author takes into account the "text as a whole" rather than only the part of it that has already been written. One encounters a situation in which the generation process depends not only on the past but also on the future, that is, on the yet unwritten part of the text. In a conversation, Arapov expressed the opinion that Zipf's law permits one to detect only the genre of a text rather than the style of a specific author.

It is of interest to study statistical dependencies in a language (in a frequency dictionary) on the basis of a new approach suggested in the present paper as well as in other recent papers of one of the authors, using the possibilities offered by modern electronic technologies. The formulas derived in [6]–[8] describe the relationship between word frequency and word rank in a frequency dictionary much more precisely than Zipf's law. Apparently, the dependencies between word frequencies and

0001-4346/2006/8056-0679 ©2006 Springer Science+Business Media, Inc.

other parameters of the dictionary may serve as an essential characteristic, if not a marker, of a writer's language.

To each word a frequency dictionary assigns its occurrence frequency (i.e., the number of occurrences) of this word in the corresponding corpus of texts. There may be several words with same frequency.

An analysis of frequency dictionaries shows that words fall into several categories:

1. "Auxiliary words" (super-frequent words, or so-called stopwords).
2. Frequent words.
3. Rare words.
4. Very rare words (condensate).

The first category includes frequent auxiliary words (prepositions, articles, and conjunctions) and pronouns. In computer science, they are referred to as stopwords and excluded from consideration. There can be large distances between frequencies in this part of the dictionary.

The second category includes words of rather high frequency. In this frequency range, the frequency spectrum typically has gaps, i.e., does not contain all the successive frequencies.

Adapted texts, where rare words are replaced by more frequent synonyms or generic terms, usually consist of words of this category.

The third category consists of mid-frequent and rare words. There are no frequency gaps in this part of the dictionary, and each frequency is shared by quite a few words.

The fourth category consists of the rarest words whose frequency lies below a given threshold.

Zipf's law is usually written in logarithmic coordinates,

$$\ln r + \frac{1}{D}\ln \omega_r = \text{const}, \tag{1}$$

where r is the rank of a word (i.e., its number in the word list arranged in descending order of frequencies), ω_r is the word frequency (i.e., the number of occurrences of the word in the text), and D is a constant, which is usually equal to 1 for dictionaries. This formula implies that rank times frequency is (approximately) a constant.

The graph of this dependence is a straight line.

Zipf's formula (1) in logarithmic coordinates gives too coarse a representation of the rank–frequency dependence; it is asymptotically true in the logarithmic variables but wrong in the original variables, without the logarithms.

Experimental and theoretical data are never compared in logarithmic coordinates in the physics literature. The reason seems to be as follows. The "averaged" curve approximating a set of experimental points gives something like an arithmetic mean for neighboring points joined by links of a broken line. The areas of the segments over and under the curve are the same.

The arithmetic mean

$$\frac{1}{n}\sum_{i=1}^{n}\ln x_i = \ln\sqrt[n]{\prod_{i=1}^{n} x_i}$$

in the logarithmic coordinates coincides with the geometric mean in the conventional coordinates, a kind of averaging that experimenters simply fail to accept.

Apart from being too coarse, Zipf's law does not describe the rare-word part of the dictionary.

Our approach differs from that adopted in linguistic statistics and is as follows. One usually treats word frequency as the probability of the word occurring in the text. We consider the problem from the opposite viewpoint. Consider an alphabetic dictionary in which word frequencies are given for each word. If we randomly pick words from this dictionary, what is the probability of picking a word with a given frequency? For example, take a word from an alphabetic list; what is the probability that this word has a prescribed frequency? This probability is equal to the number of words with this frequency divided by the total number of words in the dictionary (the dictionary length N).

Indeed, the probability of randomly picking the word with the highest frequency, say, the word *the*, is very small (it is equal to $1/N$), while the probability of coming across a word with frequency 1 is highest, because such words comprise the largest fraction of the dictionary.

In our approach, frequency plays the role of a random variable, and the number of words with a given frequency is treated as the number of outcomes favoring this frequency value. Thus, we "invert" both the frequency dictionary itself (ordered in ascending order of frequency) and the relation between a random variable and the numbers of its various outcomes. We speak of the probability of a frequency in a frequency dictionary (i.e., of a word frequency in the corresponding corpus of texts).

Thus, we consider frequencies (numbers of occurrences of words in a text) and the numbers of words corresponding to these frequencies. First, note that words with equal numbers of occurrences can be arbitrarily permuted in the dictionary (say, arranged in alphabetic or inverse alphabetic order), which clearly does not affect the relationship between frequencies and numbers of words with these frequencies. This suggests that there is an analogy with the distribution of Bose particles over energy levels ε_i.

One of the authors proved [6] that the Bose distribution can be obtained if one assumes that all distributions $\{n_i\}$ satisfying the condition

$$\sum_{i=1}^{s} n_i \varepsilon_i \le E,$$

where n_i is the number of particles on the level ε_i and E is the given total energy of all particles are equiprobable (interchangeable).

In our case, the role of the total energy of all particles should seemingly be played by the length of the corpus used to compile the dictionary (truncated by us).

However, this is not the case. As shown below on the basis of linguistic considerations, one should take into account another corpus, which we call virtual.

Indeed, to the real text of each work of fiction there corresponds a wider, more detailed virtual text.

Language gives one the opportunity of substantially sparing its material resources. There are at least two well-studied mechanisms for this, the replacement of fully meaningful text elements (such as words, phrases, and even sentences) by a limited variety of substitute words (pronouns), and ellipsis (the omission of a word that can be readily recovered). Consider, for example, the sentence "Willows grew there and birches here." The pronouns *there* and *here* refer to some phrases occurring earlier in the text, and the predicate *grew* has disappeared by ellipsis from the second part of this sentence. Thus, hidden behind this short sentence is a longer, complete sentence. One can readily imagine how much longer the text would be if these mechanisms were not used.

Using a mathematical text by way of example, we can also show how one spares language tools and reduces text length. When introducing some term for the first time, the author explains it, defines it, and introduces the corresponding notation in the form of a letter. The subsequent text uses this symbol and thereby is considerably reduced in length, since a repeated detailed description is not needed.

In encyclopedias, one makes the text shorter by denoting repeated words by their starting letters (by analogy with mathematical symbols) regardless of the grammatical form in which these words are used. There are also other examples of language effort sparing tools.

A text in which sparing mechanisms are used to a minimal extent can be very useful for a foreigner who is not fluent in the language. In computer techniques, anaphora resolution (i.e., replacing the pronouns by what they refer to and filling in the ellipses) is a necessary intermediate stage in text analysis. Without this stage, one cannot ensure the recognition of the correct semantic structure of the input text, since the computer cannot establish well-defined relations between pronouns and their antecedents. For example, [1] the text fragment "A young lady with a chaperon entered the room. She was tall." should be rewritten as "A young lady accompanied by a diminutive chaperon entered the room. The young lady was tall.", since another noun, *chaperon*, in the first sentence can potentially serve as the antecedent of the pronoun *she*.

The computer can exactly compute the length of such a "reconstructed" text and compare the actual and "reconstructed" word frequencies.

Behind each text, there are not only "spared" elements, but also implicit background knowledge shared by the writer and the potential readers. This "hidden" content permits innuendos, implications, allusions, etc. The terseness in a writer's style is highly valued by literary critics.

To connoisseurs, some well-known phrases and even single words from famous books mean whole situations, scenes, types, characters, destinies, and so on, and it suffices to communicate these isolated text fragments to reconstruct the complete picture.

All this taken together explains our thesis that the text length can be considered from the real as well as the virtual point of view. In contrast to the real length, the virtual length additionally takes into account all reconstructed substitutions and ellipses; i.e., the pronouns are replaced by their antecedents, and the ellipses are filled in by all the missing elements.

As was already mentioned, we do not take into account super-frequent words, i.e., auxiliary parts of speech, including pronouns. One should have in mind that, the more often a word occurs in a text, the more often it is replaced by a pronoun; accordingly, the more the virtual frequency of such a word is affected by these substitutions.

Thus, the increase in the frequency of a word in our virtual text is larger for larger original frequencies. We do not know how exactly larger and hence assume that to each number n_i of words there corresponds a frequency larger than i, namely,

$$\widetilde{\omega}_i = i(1 + \alpha i^\gamma), \qquad \alpha > 0, \quad \gamma > 0. \tag{2}$$

This results in a monotone decrease in the dependence of ω_i on n_i with increasing i.

We shall assume that all sets of words corresponding to the frequencies (2) in the truncated dictionary and such that the corpus comprised by these words does not exceed some virtual corpus $\mathscr{E}$ are equivalent (interchangeable).

Let $\omega_i = 1, \dots, k$ be all distinct frequencies in the frequency dictionary, and let n_i be the number of distinct words with the same frequency ω_i; then the real text length M is equal to

$$\sum_{i=1}^{k} n_i \omega_i = M,$$

and $N = \sum n_i$ is the number of words in the dictionary.

Our axiom says that all possible finite sequences $\{n_i\}$ satisfying the condition

$$\sum_{i=1}^{k} n_i \widetilde{\omega}_i = \mathscr{E},$$

where $\widetilde{\omega}_i$ is the virtual word frequency and $\mathscr{E}$ is a certain virtual text length, are equivalent (interchangeable).

Then we can use Theorem 1 in [8]. Let us number the words starting from the least frequency in the truncated dictionary and refer to the number of the word in this list as its *rank* r.

The number of *distinct* frequencies in the truncated dictionary is much less than the number of words corresponding to the dictionary.

Consider the notion of "cumulative probability" known in probability theory. It is introduced there only if an order relation is established (say, in order statistics), since it substantially depends on the order relation.

[1] *Editor's note.* The example was altered owing to differences between Russian and English grammar.

The cumulative probability $\mathscr{P}_s$ is the sum of the first s probabilities in the sequence ω_i,

$$\mathscr{P}_s = \frac{1}{N}\sum_{i=1}^{s} n_i,$$

where $s < k$.

We introduce the constant $s = \omega_0/N$ and vary the text length for each writer, say, by truncating the text on some page and then by adding k pages. Let us plot the curves showing the numbers of discarded words with maximum and minimum frequency against N. The parameter s computed from the least-square deviation from the logarithmic law is also a function of N. To some extent, these curves might be markers of the literary work, the author, or at least the genre.

This means that $r_s = N\mathscr{P}_s$ is the rank (number) of the word if we successively count all words starting from $\omega_{\min}$, the order of words corresponding to the same frequency being irrelevant.

By virtue of our axiom, the rank obeys the following formula ([6, Theorem 1]; see also [8]):

$$r_l = \sum_{i=1}^{l} \frac{c}{e^{\beta\widetilde{\omega}_i} - 1}, \tag{3}$$

where r_l is the rank of the lth word. Since $k \ll N$, it follows that $\beta \ll 1$, and we obtain

$$r_l = \frac{1}{\beta}\sum_{i=1}^{l} \frac{1}{\omega_i + \alpha\omega_i^{1+\gamma}} + O\biggl(\frac{1}{\beta^2}\biggr)$$

as $\beta \to 0$ and $\sigma = 0$. Since, by [6, Theorem 1], $l > \varepsilon k$, where ε is arbitrarily small but independent of $k \to \infty$, and $\Delta\omega_i = 1$, it follows that

$$r_l \cong \sum_{i=1}^{l} \frac{1}{\omega_i + \alpha\omega_i^2} = \int^{\omega} \frac{d\omega}{\omega + \alpha\omega^{1+\gamma}} = \ln\frac{\omega^\gamma}{1 + \alpha\omega^\gamma} + c$$

as $N \to \infty$, and hence

$$r_l = T\,\ln\varrho\frac{\omega_l^\gamma}{1 + \alpha\omega_l^\gamma}, \qquad T = \frac{1}{\beta}, \tag{4}$$

where $\varrho = \exp(c/T)$.

The two constants occurring in the formula in [6, Theorem 1], as well as the corresponding constants in the formula for the Bose–Einstein distribution, are defined via the total number of particles (i.e., the length of the dictionary in our case) and the total energy (i.e., the virtual text length).

We do not know the virtual text length and do not know exactly where to truncate the dictionary. Hence we have to find a different way to determine the four constants occurring in (4).

First, note that $\gamma \sim 1$ for dictionaries. (This is already known from Zipf's studies.) Hence we set $\gamma = 1$ for now. Then we choose two points, one of small rank and the other of medium rank, where the theoretical curve should exactly fit the experimental points. These points should be stable in the sense that the theoretical curve does not change to within the prescribed accuracy if we replace these points by nearby points.

To find the last constant, we must minimize the deviation of the theoretical curve from the experimental points. This permits one to determine the truncated dictionary length N as well as the virtual text length $\mathscr{E}$.

Next, setting

$$r_c = N - r,$$

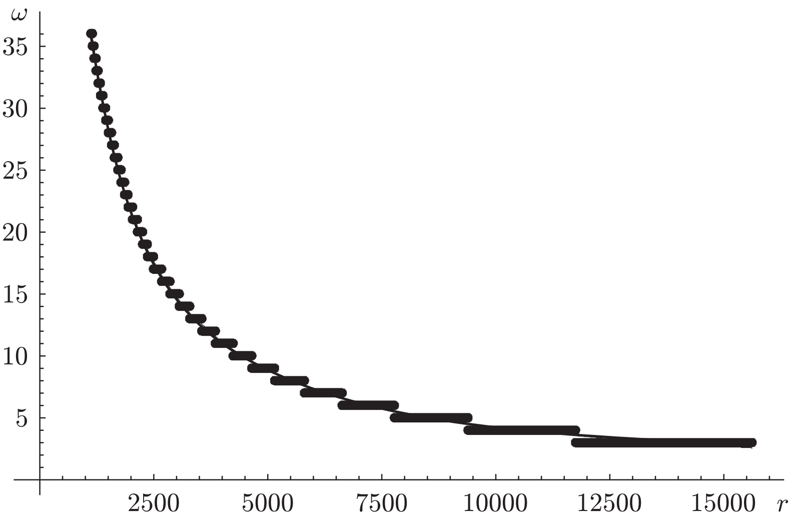

Fig. 1. The variable r is the word's number in the frequency dictionary in ascending order of frequencies; ω is the word frequency

where r_c is the word rank counted from the highest frequency $\omega_{\max}$, we arrive at the relation

$$\frac{r_c}{r_c^0} = \ln\biggl(1+\alpha\biggl(\frac{1}{\omega}\biggr)^{\gamma}\biggr) + c_1, \tag{5}$$

where r_c^0 is the normalization constant. As $\omega \to \infty$, this relation becomes Zipf's law

$$\ln r_c + \gamma \ln \omega \sim \text{const}\,.$$

From this, using available software, we find γ. As a rule, $\gamma = 1$ for frequency dictionaries. Figures 1 and 2 show the rank–frequency curves for the first volume of Leo Tolstoy's *War and Peace.* In Fig. 1, the low-frequency part of the dictionary is approximated by formula (5) (words with frequencies $\geq$ 3; $r_0 = 43408.8$, $\alpha = 0.891995$, and $c_1 = -189.321$). For the entire dictionary, Fig. 2 shows the deviation of the theoretical data from the dictionary data (the difference between the frequency given in the dictionary and the frequency given by the formula) against rank. The deviation is seen not to exceed 1.5 words if the rank is larger than 300.

After that, one can repeat the entire procedure and find the virtual length $\mathscr{E}$ more exactly.

We introduce the parameter l defined as the text length counted from the beginning and obtain the curves $\mathscr{E}(l)$, $N(l)$, and $\omega_{\min}(l)$, which are markers of a given literary work.

2. SEMIOTICS OF AN UNORDERED SET OF SIGNS

The logarithmic law derived by one of the authors [6], [8], which substantially refines and simultaneously proves Zipf's law for frequency dictionaries, can be extended to a wide class of semiotic systems, for example, to city population statistics, whose simplest version resembling Zipf's law was known earlier. The Lotka and Bradford laws in documentation science, which were earlier derived empirically [9], are also refined and covered by this logarithmic law.

We generalize this law to semiotic objects by considering arbitrary signs and intensity of their usage. This generalization will in particular imply a refinement of the Lotka and Bradford laws.

The main property of a sign is its recurrence in a given socium. A sign used by nobody but the author is not a sign. The abstruse words invented by Khlebnikov can serve as an example of "nonsigns."

There are various types of semiotic objects (signs).

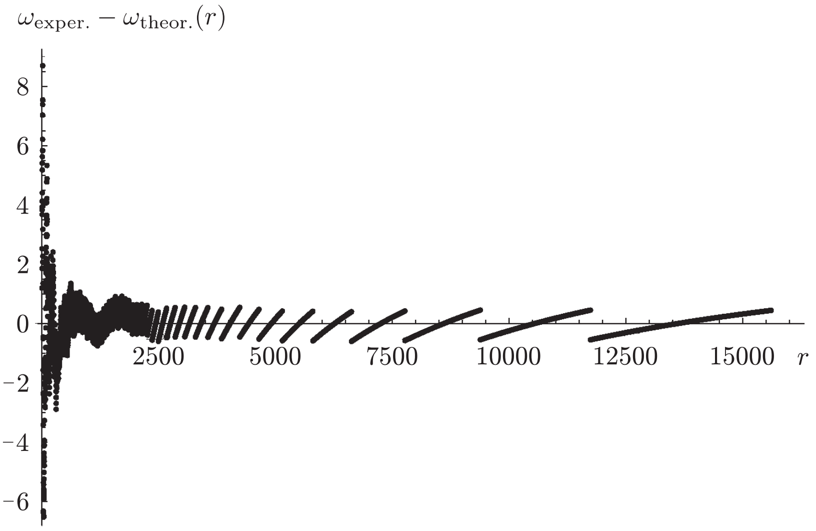

Fig. 2. The variable $\omega_{\text{exper.}}$ is the real word frequency in the frequency dictionary compiled for the first volume of *War and Peace*

Letters, as well as words and hieroglyphs, are signs of natural language. If we consider a writer's work, then we deal with a united text and its vocabulary. The vocabulary is the set of distinct words occurring in the text. Here the text is treated as a linear sequence of syntagmatically related words. If a heterogeneous corpus is considered, then the text is a set of words rather than a linear sequence.

A text in a natural language has a real length and a virtual length. This is related to the economy principle inherent to the language. For example, as we mentioned above, text elements (words or even sequences) that are repeated or known to the recipient (reader) are replaced by a limited variety of substitute words (pronouns), and easy-to-figure-out words and constructions are omitted altogether. (The latter mechanism, as we mentione above, is known as *ellipsis.*) Thus, there is usually a longer text "hidden" in a short text. One can readily grasp how longer this virtual text is by trying to reconstruct all omitted, substituted, or implied text elements. In particular, this procedure is a necessary stage in automated (computer) text analysis. Obviously, the virtual text length is substantially increased owing to background knowledge shared by the author and the reader and taken into account by the author when writing the text.

Following Humboldt and Prieto, we refer to this virtual length as *energeia*[2] and denote it by $\mathscr{E}$. This concept becomes increasingly topical in modern linguistics. We shall shortly show that the notions of virtual reality and energeia also apply to other sign objects.

Let us introduce some terminology concerning semiotic objects and their properties. First, we apply these terms to natural language, which is the best studied and most comprehensible type of sign systems.

A word of natural language is a *sign.* The set of words is a *sign dictionary.* The number of times a sign is used characterizes the intensity of usage of that sign. This characteristic will be called the *real cardinality* of the sign. In natural language, this is the word's frequency in texts. The *real cardinality of a sign dictionary* is the total number of occurrences of all signs in the dictionary. In language systems, the cardinality of the dictionary is related to the corpus used to compile the dictionary (set of words).

[2] Starting from the time of the famous German linguist von Humboldt, the Greek word $\varepsilon\nu\varepsilon\rho\gamma\varepsilon\iota\alpha$ (creative force) is used in semiotics to refer to the subjective structure of the realm of a work of fiction; this structure operates, is preserved, and reproduces itself, i.e., resides in "motion" [10, pp. 372–399].

A *sign* also has a *virtual cardinality.*[3] The real cardinality of a word is the number of its explicit occurrences in a given corpus. The virtual cardinality counts all occurrences, even those where the word is not used explicitly but is necessary for the correct analysis and perception of the text by the reader. This includes the cases in which the word is replaced by a pronoun, omitted (ellipsis), or hidden in the implication. The virtual cardinality of a word exceeds the real cardinality and increases with the latter.

Some signs have equal real cardinality (words with equal frequency).

Likewise, the sign dictionary is characterized by the *virtual cardinality*, or energeia, $\mathscr{E}$.

Jumping ahead, note that we are interested in the number of signs whose cardinality is greater than or equal to an *a priori* given cardinality, or those whose cardinality is less than an *a priori* given cardinality. For a frequency dictionary, the number of words with frequency greater than a given frequency is the rank, counted in descending order of frequencies, of a word corresponding to that frequency; accordingly, the number of words with frequency smaller than a given frequency is the word rank counted in ascending order of frequencies. Accordingly, we introduce the notion of rank for signs in sign dictionary.

Let us give examples of other sign systems.

During elections, a nominated candidate or party is a sign. The list of candidates is a sign dictionary. The number of votes in favor of a candidate corresponds to the real cardinality of the sign. The virtual cardinality is the number of favorable votes plus the number of people who support the candidate but do not vote for some reason (the candidate's rating).

A city depicted on a map by its name in bold or fine print and by a circle of certain thickness and diameter is a sign indicating that city's population. The number of registered residents of the city is the real cardinality of the sign. The total number of people in the city is its virtual cardinality. The set of all cities in a country is a city dictionary. The real cardinality of the sign (city) dictionary is the country's population; the virtual cardinality is the total number of people staying there, including temporary workers, tourists, etc. (the energeia $\mathscr{E}$ of the country).

A person's name is a sign; the real cardinality of a name is the statistical data on the number of people with this name in a given region. The virtual cardinality of the name is the number of all people with this name who live in the region, regardless of whether they are registered or not. The virtual cardinality can also involve foreigners who use customary local names instead of their exotic foreign names.

In a library, a book with a given title is a sign. The book catalog is a sign dictionary. The number of requests for a given book is the real cardinality of the sign (the demand for, or the popularity of, the book). The total number of requests for all books is the real cardinality of the sign dictionary. The virtual cardinality of a book treated as a sign includes all instances of usage the book, not only by the reader who himself borrowed the book from the library but also by his friends, neighbors, and relatives who have also read it.

An internet site is also a sign. The site catalog is the sign dictionary. The real cardinality of a site is the registered number of visits to the site by various users. The virtual cardinality of a site is the number of actual visits plus the usage of site materials without visiting the site.

An icon (e.g., of St. Nicolas) is a sign recognizable in various realizations. The collection of icons in a church is a sign (icon) dictionary. The real cardinality of an icon treated as a sign is the number of parishioners appealing to the icon, which is registered by the number of candles lit before the icon (assuming that nobody lights more than one candle). The virtual cardinality of an icon includes all mental appeals to the icon by the parishioners.

A church, in turn, is a sign as well. The sign dictionary is the set of churches in the city (forty forties[4]). The real cardinality of a church is the number of candles lit in front of all icons. The

[3] Duns Scot (followed by Suarez) was apparently the first to notice, as early as in the 13th century, that Aristotle's *actus* and *potentia* should be supplemented with *actus virtualis*, a most important intermediate notion.

[4] *Translator's note.* The proverbial number of churches in Moscow.

virtual cardinality of a church is the number of people that come to pray. This example shows that signs may be hierarchical.

A museum can also be viewed as a sign. The sign dictionary is the list of all the museums in a city. The number of tickets sold is the real cardinality. The virtual cardinality is the number of people who go to the museum, including those who do not attend the exposition owing to high ticket price or lack of time or restrict themselves to leafing through the catalog.

Consider an advertising campaign with some list of goods (sign dictionary). The real advertisement cardinality for an individual article treated as a sign is the number of people who have taken the advertising pamphlet. The virtual advertisement cardinality is the total number of people who have looked through the pamphlet.

Let us proceed to the main formula generalizing the formula derived in [8] for frequency dictionaries to a wider class of semiotic objects.

Consider an abstract sign dictionary $\{s_i\}$, $i = 1, 2, \dots, n$. Let the real cardinality of the sign s_i be equal to ω_i.

We introduce two unknown constants α and γ and assume that the virtual cardinality of the sign s_i is equal to

$$\omega_{\text{vir}} = \widetilde{\omega}_i = \omega_i(1 + \alpha\omega_i^\gamma). \tag{6}$$

If the signs are words (instructions) of a simple programming language, then the virtual length of a software package may prove to be equal to the real length, so that $\gamma = 0$.

Our paradigm is based on Kolmogorov's definition of randomness as maximal complexity (Kolmogorov complexity) [11]. This means that, the longer is the algorithm used by the author to invent a linear sequence (syntagmatic set) of signs, the closer he is to the general position of a majority of all possible sequences. Likewise, when playing heads and tails, the longer and more complicated the player's algorithm, the closer the outcomes to the generic case in which heads, as well as tails, fall half of the times.

If a set of signs (e.g., the distribution of population over cities or the visits to cites) was formed randomly (i.e., by a long algorithm if one keeps track of its deterministic history), then, as a rule, it will be in general position, i.e., near the points where a majority of possible outcomes is concentrated.

Let n_i be the number of distinct signs of equal real cardinality ω_i.

Since

$$\sum_{i=1}^{k} n_i\widetilde{\omega}_i = \mathscr{E},$$

we assume that the number of signs n_i corresponding to a given virtual cardinality ω_i of the sign s_i is a random variable uniformly distributed over all sets $\{n_i\}$ such that

$$\sum_{i=1}^{k} n_i\widetilde{\omega}_i \le \mathscr{E}.$$

In other words, *the sets* $\{n_i\}$ *are elementary events.* This is our main (and unique) axiom. Obviously, $\mathscr{E} \le \widetilde{\omega}_{\max}N$, where N is the sign dictionary length.

The problem splits into two cases:

1. $\mathscr{E} < \dfrac{\sum_{i=1}^{k}\widetilde{\omega}_i}{k}N$.
2. $\dfrac{N}{k}\displaystyle\sum_{i=1}^{k}\widetilde{\omega}_i \le \mathscr{E} \le \widetilde{\omega}_{\max}N$.

This axiom should be understood as follows. Our specific sequence of signs with "energeia" $\mathscr{E}$ is one of the numerous similar sequences with energeia not exceeding $\mathscr{E}$ and with the same sign dictionary, and we assume that, at least for the main part of the signs, it is in general position with respect to all possible $\{n_i\}$ satisfying condition 1 or 2.

Let us number the signs in the dictionary in ascending order of cardinality, starting from the minimum cardinality $\omega_{\min}$. Signs of equal cardinality can be ordered arbitrary. The number of a sign under this ordering will be called the *rank* of that sign and denoted by r. If l is the number of a cardinality ω_l (starting from the minimum cardinality $\omega_{\min}$), then by r_l we denote the number of all signs whose cardinality does not exceed ω_l.

By our axiom, in case 1 we can apply the formula in [6, Theorem 1] (see also [8]) to the sign rank:

$$r_l = \sum_{i=1}^{l} \frac{c}{e^{\beta\widetilde{\omega}_i} - 1}, \tag{7}$$

where r_l is the rank of the lth sign. Since $k \ll N$, it follows that $\beta \ll 1$; from this, for $\sigma = 0$, we obtain formula (4) for the sign rank as $\beta \to 0$.

In a similar way, one can prove that

$$r_l = \sum_{i=1}^{l} \frac{c}{1 - e^{-\beta\omega_i}} \tag{8}$$

in case 2, which results in the same relation (4) as $\beta \to 0$.

The two constants occurring in the formula in [6, Theorem 1], as well as the corresponding constants in the formula for the Bose–Einstein distribution, are defined via the total number of particles (here this corresponds to the sign dictionary length) and the total energy (the virtual length of the sign dictionary).

We do not know the virtual length of the sign dictionary and do not know where exactly to truncate the dictionary. Hence we have to find a different way to determine the four constants occurring in (4).

First, we determine γ from Zipf's law. Then we choose two points, one of small rank and the other of medium rank, where the theoretical curve should exactly fit the experimental points. These points should be stable in the sense that the theoretical curve does not change to within prescribed accuracy if we replace these points by nearby points.

To find the last constant, one should minimize the deviation of the theoretical curve from the experimental points. This permits one to determine the truncated dictionary length N as well as the virtual text length $\mathscr{E}$.

Next, setting

$$r_c = N - r,$$

where r_c is the sign rank (number) counted from the highest cardinality $\omega_{\max}$, we arrive at the relation

$$\frac{r_c}{r_c^0} = \ln\biggl(1 + \Bigl(\frac{\omega_0}{\omega}\Bigr)^{\gamma}\biggr),$$

where ω_0 and r_c^0 are the normalization constants. As $\omega \to \infty$, this relation becomes Zipf's law

$$\ln r_c + \gamma \ln \omega \sim \text{const}\,.$$

From this, using available software, we refine γ. After that, one can repeat the entire procedure and find the virtual length $\mathscr{E}$ more exactly.

In conclusion, we give two important examples of signs and their cardinalities, for which a sufficiently accurate statistics is available.

The first example (see Fig. 3) deals with American car brands (as signs) and their prices (as the cardinalities of the corresponding signs).

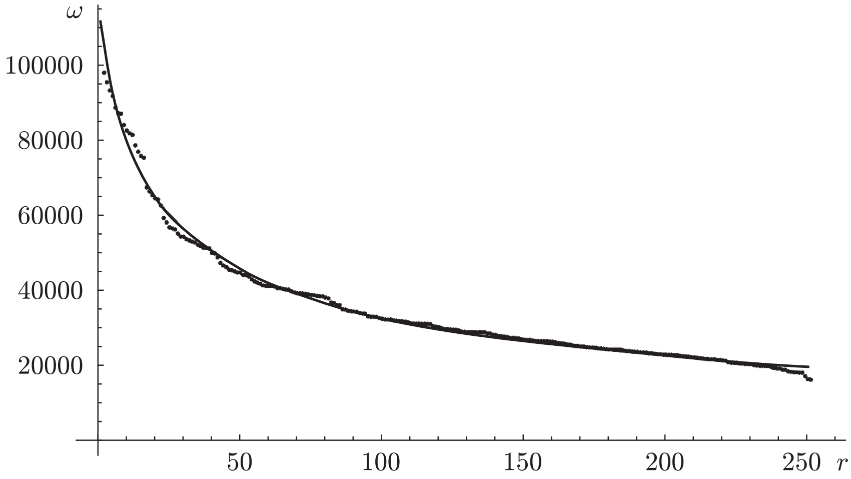

Fig. 3. American car brands and prices

The virtual cardinality includes not only the car's price, but also a plethora of additional expenses, including insurance, gas, and parking costs as well as services taken into account by the buyer. The rank in Fig. 3 is the brand number in descending order of the price ω.

However, this sign (the car brand corresponding to a given price) substantially differs from all the above-mentioned signs. The accompanying expenses indeed increase with increasing price starting from some ω_{cr}. (On the graph in Fig. 3, $\omega_{\mathrm{cr}} \sim 22500$.) However, the accompanying expenses also increase if the price goes below the critical level: safety declines, and the probability of an accident increases. The buyers of cheaper cars often use them until complete wear-out; old cars are subject to a heavily increased road tax (abroad), etc. Hence, in contrast to (6), we should write

$$\omega_{\mathrm{vir}} = \widetilde{\omega}_i = \omega_i(1 + \alpha\omega_i^{\gamma} + \delta\omega_i^{-\sigma}),$$

i.e., introduce two more unknown constants. Now we have

$$r \sim c_1 \int^{\omega} \frac{d\omega}{\omega(1 + \alpha\omega^{\gamma} + \delta\omega^{-\sigma})} + c_2.$$

One of the constants can be determined from the condition

$$\left(\frac{1}{\omega(1 + \alpha\omega^{\gamma} + \delta\omega^{-\sigma})}\right)' = 0$$

for $\omega = \omega_{\mathrm{cr}}$. If the "tail" below ω_{cr} is not long, then, for simplicity, one can set $\sigma = \gamma$ and $\delta = 1/\alpha$; then

$$r = \frac{c_1}{1 + \kappa\omega^{\gamma}} + c_2,$$

and hence

$$\omega = \mathrm{const}\left(\frac{r}{r_c}\right)^{1/\gamma},$$

which simplifies fitting the coefficients κ and γ.

Price variations for a financial instrument (share) at stock exchange can be conveniently denoted by so-called "Japanese candles.' Japanese candles, joined into classes with accuracy of 0.2%, are an example of a sign system. We number them in descending order of occurrence frequency. The occurrence frequency of a Japanese candle sign is its cardinality. Its virtual cardinality is related to deals concluded by brokers outside the exchange in electronic trade networks and similar financial

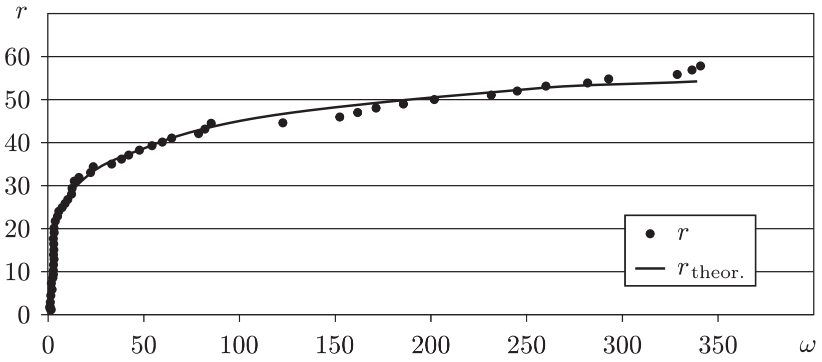

Fig. 4

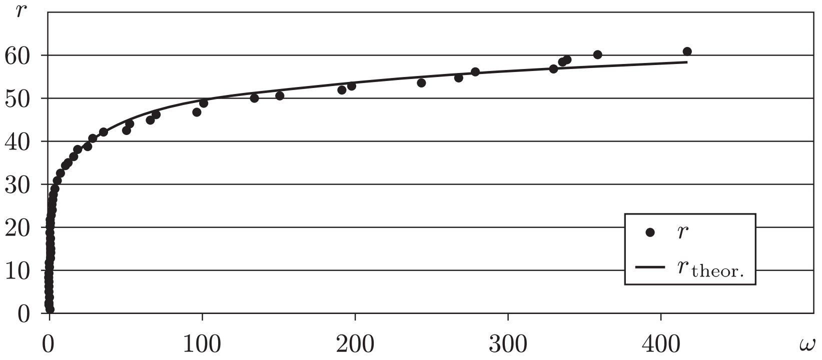

Fig. 5. The Japanese candle order number vs. frequency: r is the candle order number, starting from the smallest candle; ω is the candle frequency

instruments that can replace the given financial instrument for the investor. Recently there has been a tendency towards grouping shares into equivalence sets (by analogy with grouping words into descriptors). For an investor, the shares within a group are interchangeable, just as pronouns or rare synonyms can replace repeated words in a text and hence affect the word occurrence statistics. Investors avoid recurrence even more than writers avoid repetitions of the same words, for the risk of losing decreases with increasing range of purchases. The rank in Figs. 4 and 5 is the candle number, starting from the smallest candle.

ACKNOWLEDGMENTS

The authors are deeply grateful to M. Yu. Romanovskii for data about car prices in the USA in 2005, to A. V. Churkin for help in processing these data, and to N. V. Starchenko for providing and helping to process financial instrument data.

This research was supported by the Russian Foundation for Basic Research under grant no. 05-01-00824.

BIBLIOGRAPHY

1. S. M. Gusein-Zade, “On the frequency distribution of Russian letters,” *Problemy Peredachi Informatsii* [*Problems Inform. Transmission*], **23** (1988), no. 4, 102–107.
2. M. B. Malyutov, “A survey of methods and examples of text atribution,” *Obozrenie Prikladnoi i Promyshlennoi Matematiki*, **12** (2005), no. 1, 41–77.

3. N. A. Morozov, "Linguistic spectra: a tool for distinguishing plagiarisms from true works of known authors. A stylemetric study," *Izv. Otd. Russ. Yazyka i Slovesnosti Imper. Akad. Nauk*, **XX** (1915), no. 4.
4. A. A. Markov, "On an application of the statistical method," *Izv. Imper. Akad. Nauk*, *Ser. VI*, **X** (1916), no. 4, 239.
5. M. V. Arapov, E. N. Efimova, and Yu. A. Shreider, "On the meaning of rank distributions," *Nauchn. Tekhn. Inform.*, *Ser.* 2 (1975), no. 1, 9–20; `http://kudrinbi.ru/public/442/index.htm`.
6. V. P. Maslov, "On a general theorem of set theory resulting in the Gibbs, Bose–Einstein, and Pareto distributions and the Zipf–Mandelbrot law for stock market," *Mat. Zametki* [*Math. Notes*], **78** (2005), no. 6, 870–877.
7. V. P. Maslov, "Linguistic statistics," *Russ. J. Math. Phys.*, **13** (2006), no. 3, 315–325.
8. V. P. Maslov, "The lack-of-preference law and the corresponding distributions in frequency probability theory," *Mat. Zametki* [*Math. Notes*], **80** (2006), no. 2, 220–230.
9. S. P. Kapitsa, S. P. Kurdyumov, and G. G. Malinetskii, *Synergetics and Forecasts of the Future* [in Russian], Editorial URSS, Moscow, 2001.
10. *Semiotics*: *Semiotics of Language and Literature* [in Russian], Raduga, Moscow, 1983.
11. A. N. Kolmogorov, "Theory of information transmission," in: *Works of Session of Akad. Nauk SSSR on Production Automation* [in Russian], Moscow, 1957.

V. P. Maslov
Moscow State University
E-mail: `v.p.maslov@mail.ru`

T. V. Maslova
Moscow Institute of Economics