Phoneme Frequencies Follow a Yule Distribution

(The Form of the Phonemic Distribution in World Languages)

Yuri Tambovtsev and Colin Martindale

Frequency of occurrence of words in a language is well described by Zipf's (1949) law. However, Zipf's law does not well describe the distribution of the phonemes from which words are composed. Examination of frequency of occurrence in 95 languages shows that phoneme frequencies are best described by an equation first developed by Yule (1924) that also describes the distribution of DNA codons. The Yule equation fits the distribution of phoneme frequencies better than the Zipf equation or equations proposed by Sigurd (1968) and Borodovsky and Gusein-Zade (1989).

Keywords: phoneme frequency, phoneme distribution, cross-linguistic research,

There are about 6,000 languages and dialects in the world. They are divided into several language families according to their genetic origin. We have taken for our study the data of the frequency of occurrence of phonemes in the following 95 languages: 1) Indo-European; 2) Finno-Ugric; 3) Samoyedic; 4) Turkic; 5) Mongolic; 6) Tungus-Manchurian; 7) Yeniseian; 8) Caucasian; 9) Paleo-Asiatic; 10) Sino-Tibetan; 11) Afroasian; 12) Bantu; 13) Austro-Asiatic; 14) Australian and 15) American Indian. The exact values of the frequency of phonemic occurrence may be found elsewhere (Tambovtsev, 2001-a; 2001-b; 2001-c; 2003) . The goal of the present paper is to investigate the form of the distribution of the frequency of the occurrence of phonemes in the languages mentioned above. It is interesting to see if there is any difference in the form of distribution according to the languages.

The frequency of occurrence of words in a language is well described by Zipf's law:

$$(1) \qquad F_r = \frac{a}{r^b},$$

where F_r is the frequency of the word ranked r, r is the rank of the word when frequencies are ranked from most frequent (rank = 1) to least frequent (rank = n), and a and b are parameters to be estimated from the data. The usual finding is that b is close to 1 (Zipf 1949; Kučera and Francis 1968). When Zipf's law is represented in terms of logs the result, graphed on log-log coordinates, is a straight line with a slope of about -1. Zipf's law is not restricted to language. It also seems to describe a number of phenomena that are distributed in a very skewed way ranging from wealth (Pareto 1896), through number of publications by scientists in a discipline (Lotka 1926), to the size of cities in a country (Simon 1955). Zipf's law does not describe all highly skewed distributions. For example, it provides a poor description of the distribution of the eminence of poets (Martindale 1995). Zipf's law does not describe the frequency distribution of DNA codons (Borodovsky & Gusein-Zade, 1989; Martindale & Konopka 1996) or the frequency distribution of phonemes or graphemes in a language very well (Witten & Bell 1990; Martindale,

Gusein-Zade, McKenzie & Borodovsky 1996). In all of these cases, it usually overestimates both high-frequency and low-frequency items and underestimates items of moderate frequency.

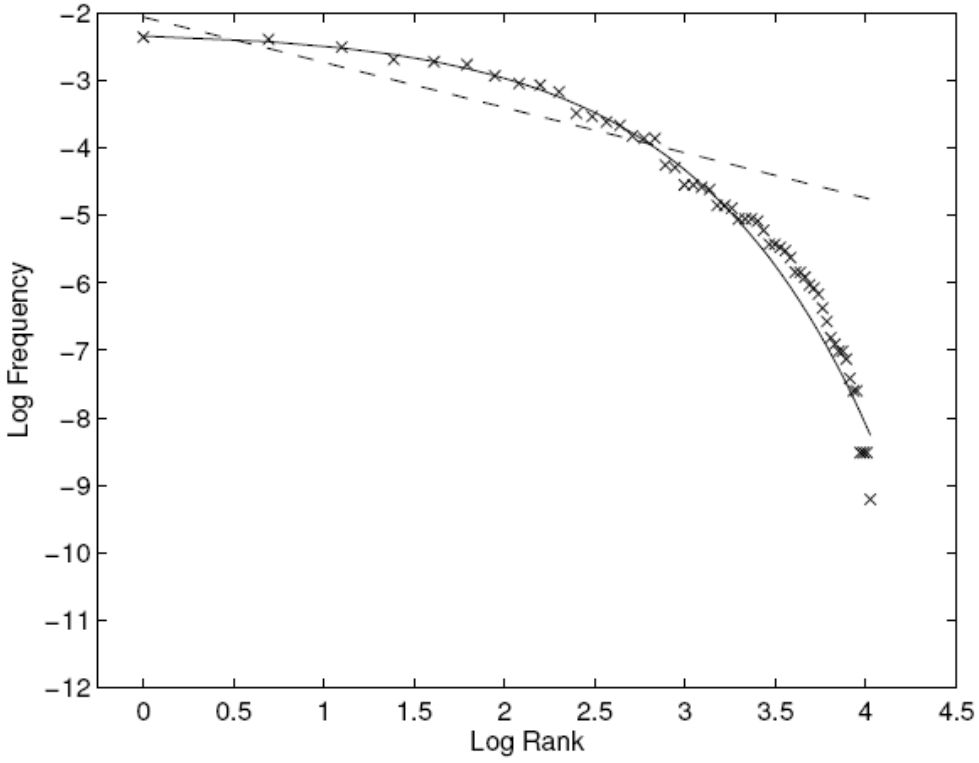

Figure 1 - *Fit of Zipf (----) and Yule (—) equations to the ranked frequency distribution of Finnish phonemes*

An example for Finnish phonemes is shown in Figure 1. In cases where the Zipf equation does not fit a ranked frequency distribution, a related equation first proposed by Yule (1924) to describe the number of species per genus,

(2) $$F_r = \frac{a}{r^b} c^r$$

often corrects the problem (Simon 1955). Note that the Zipf equation is a special case of the Yule equation in which c^r is neglected. It is not always possible to neglect this term. As shown in Figure 1, the Yule equation fits the distribution of phoneme frequencies in Finnish much better than the Zipf equation. As we show below, this is not an isolated case. Pareto (1898) wrote before Zipf or Yule and used a different equation. His data on wealth and income can be recast

so that they are well fit by a Zipf equation. He noted, however, that extremely low incomes (compare the low-frequency phonemes to the right of Figure 1) are removed from the population as a person with such an income would starve to death or income is brought up to a subsistence level by welfare. Were this not the case, income and wealth would probably be distributed in a fashion better described by a Yule than a Zipf distribution. There is no welfare mechanism to rescue poets from obscurity, so poetic eminence is better described by a Yule than a Zipf distribution. Yule distributions are probably far more common than realized. In studies purporting to find Zipf distributions, it is often rather clear that extremely infrequent events or small entities are simply being ignored. This can be seen in Zipf's (1949) own work.

The data shown in Figure 1 suggest that it might be the case that F_r, rather than the log of F_r is dependent upon the log of r,

(3) $F_r = a + b \log r$

However, a similar but simpler equation has been proposed. Gusein-Zade (1987, 1988) and Borodovsky and Gusein-Zade (1989) suggested the equation,

(4) $F_r = (1/n) (\log (n + 1) - \log r)$

where n is the number of symbols. In this case, F_r – rather than the log of F_r – is predicted from the log of r. Because n is always known in advance, it is not necessary to estimate it. Thus, the equation has the merit of being parsimonious because no parameters need be estimated. Gusein-Zade (1988) presented mainly graphical evidence that the parameter-free equation fits the frequency distributions of English, Estonian, Russian, and Spanish graphemes quite well. Good (1969) had previously suggested the same equation in a different form to describe the ranked distribution of phoneme and grapheme frequencies:

(5) $$F_r = \frac{1}{n} \sum_{i=r}^{n} \frac{1}{i}.$$

Good gave data from two samples of English – one for phonemes and one for graphemes – in support of the equation.

Sigurd (1969) suggested a geometric-series equation for the ranked distribution of phoneme frequencies:

(6) $F = a \bullet k^{r-1}$,

where a is the frequency of the most frequent phoneme, k is the parameter to be estimated, and r is rank. Because Equation 6 predicts the frequency of the most frequent phoneme trivially by making the predicted frequency equal to the observed frequency, Sigurd (1968) proposed a related one-parameter equation that avoids this problem,

(7) $$F_r = \frac{(1-k)\,k^{r-1}}{1-k^n}$$

In this case, $\underline{n}$ is the number of symbols, and k is the parameter to be estimated. For the equation to work, the frequencies have to be normalized so they sum to 1.00. Equation 7 is a special case of the Yule equation (Equation 2) for which, in terms of Equation 2, $b = 0$, $c = k$, and

(8) $$a = \frac{1 - k}{k(1 - k^n)}.$$

It is thus mathematically impossible for the Sigurd equation to fit a set of data better than the Yule equation, because it has less parameters. This can be seen as an advantage though. The Sigurd equation has the merit of having only one rather than three free parameters to estimate. Sigurd (1968) found that Equation 7 better describes phoneme frequencies in five languages than the Zipf equation, but he gave no statistical measures of the goodness of fit of the two equations.

For 18 language samples, Martindale, Gusein-Zade, McKenzie, and Borodovsky (1996) compared the fit of the Borodovsky and Gusein-Zade parameter-free equation (Equation 4) with the fit of the Zipf (Equation 1), Yule (Equation 2), and Sigurd (Equation 7) equations to ranked frequency distributions of phonemes. The average fit of both the Sigurd and the Borodovsky and Gusein-Zade equation as measured by R^2 was .95–close to the maximum possible value of 1.00. The Zipf equation did not fit the data as well. The average R^2 was .88. The Yule equation provided the best fit (average $R^2 = .98$) but at a cost of estimating three parameters. Examination of residuals showed that the Zipf equation consistently overestimated both high- and low-frequency phonemes. The Yule, Sigurd, and Borodovsky and Gusein-Zade equations tend to overestimate slightly the frequency of low-frequency phonemes. Similar results were found for graphemes from 14 languages. Given the small number of cases, the authors could not demonstrate statistically that the Yule equation provides the best description of the distribution of phoneme or grapheme frequencies. They could only draw the weak conclusion that it is a matter of taste whether one prefers greater precision (the Yule equation) or for greater parsimony (the Borodovsky and Gusein-Zade equation or the Sigurd equation).

The purpose of the present paper was to examine the fit of the Zipf, Yule, Sigurd, and Borodovsky and Gusein-Zade equations to phoneme frequencies in a large number of languages. This gives us the statistical power to determine if one of these equations is better than the others in describing the distribution of phoneme frequencies. Our sample of 95 languages does not include languages from all language families. American and African languages are notably underrepresented. We were limited to languages for which we could find or compute tabulations of phoneme frequencies. This must be kept in mind in drawing conclusions. Tabulations of phoneme frequencies were taken from Altmann (1993) for Hawaiian; Dewey (1923) for American English; data for Martindale et al. (1996) for ancient Greek; Sigurd (1968) for Samoan and Kaiwa; Weiss (1961) for Swedish, and Zipf (1929) for Sanskrit. Data for all other languages come from counts made by the first author (Tambovtsev 1994-a; 1994-b; 2001-a; 2001-b; 2001-

c; 2003). The data on the frequency of occurrence of phonemes in 159 world languages have been taken from the counts of Yuri Tambovtsev who computed them on the running texts of fictional prose. The texts in Indo-European, Finno-Ugric, Turkic, Paleo-Asiatic, Afroasian (Semito-Hamitic), Sino-Tibetan, Austronesian, Australian, Austro-Asiatic and American-Indian language families were taken for the phono-statistical investigation. The style of the fictional prose has been taken for the reasons of commensurability. The details on the phonemic data in different styles can be found elsewhere (Tambovtsev 2003).

The fit of the equations as described by R^2, along with other relevant information, is given in Table 1. We have organized the language into the families suggested by Ruhlen (1987). This is only for convenience and should not be taken as meaning that we take any position concerning Ruhlen's classification system. The average R^2 values shown in the table were determined by using Fisher's *r-z* transformation (McNemar, 1962).

Language Family Language (dialect)	R^2				Phonemes	
	Zipf	Sigurd	Borodovsky & Gusein-Zade	Yule	In Language	In Sample
Altaic						
Azerbaijani	0.88	0.98	0.98	0.99	32	91,706
Buryat	0.83	0.99	0.98	0.99	39	87,751
Even (Lamut)	0.85	0.98	0.99	0.98	41	61,126
Evenk	0.87	0.98	0.98	0.99	25	138,146
Hakas	0.92	0.89	0.94	0.96	35	193,782
Jakut	0.87	0.99	1.00	1.00	40	236,245
Japanese	0.86	0.98	0.95	0.98	47	95,076
Kalmyk	0.92	0.94	0.98	0.98	41	33,240
Karakalpak	0.85	0.98	0.97	0.98	33	201,865
Kazah	0.93	0.92	0.96	0.97	32	89,517
Kirgiz	0.88	0.98	0.98	0.99	32	29,935
Mongolian	0.90	0.98	0.99	0.99	42	87,625
Nanaj	0.91	0.95	0.95	0.98	59	160,568
Negidal	0.89	0.93	0.97	0.96	38	88,437
Oroch	0.94	0.87	0.92	0.97	46	123,761
Tatar (Barada)	0.88	0.97	0.96	0.98	45	67,569
Turkish	0.90	0.97	0.99	0.98	32	107,192
Turkmen	0.84	0.95	0.94	0.95	29	32,400
Ujgur	0.89	0.97	0.98	0.98	34	48,331
Ulch	0.92	0.97	0.94	0.99	50	39,055
Uzbek	0.91	0.89	0.92	0.95	34	40,894

Indo-European

Albanian	0.82	0.98	0.96	0.98	25	76,200
Armenian	0.98	0.89	0.87	0.99	39	88,300
Bulgarian	0.83	0.99	0.98	0.99	48	18,102
Czech	0.87	0.96	0.98	0.97	37	186,641
Dutch	0.94	0.94	0.97	0.99	39	2,220,000
English (American)	0.86	0.96	0.97	0.97	42	500,000
English (British)	0.92	0.95	0.97	0.98	44	217,558
French	0.89	0.93	0.96	0.96	36	1,390,900
German	0.91	0.94	0.96	0.98	51	100,000
Greek (Modern)	0.83	0.99	0.98	0.99	25	17,690
Greek (Homeric)	0.81	0.98	0.96	0.98	24	494,315[b]
Gujarati	0.92	0.93	0.97	0.97	29	475,828
Gypsy	0.96	0.87	0.86	0.98	45	56,431
Hindi	0.92	0.97	0.97	1.00	50	507,904
Italian	0.78	0.99	0.96	0.99	49	84,098
Latin	0.79	0.97	0.95	0.97	21	351,580
Latvian	0.88	0.97	0.99	0.98	33	60,000
Lithuanian	0.90	0.94	0.98	0.99	64	20,000
Maharashtri	0.98	0.83	0.75	0.98	32	9,443
Marathi	0.94	0.92	0.96	0.98	38	236,127
Moldavian	0.95	0.88	0.93	0.96	24	50,000
Norwegian	0.90	0.94	0.97	0.97	36	146,908
Osetin	0.96	0.90	0.93	0.98	35	103,364
Persian	0.92	0.96	0.99	0.99	31	292,314
Polish	0.89	0.91	0.94	0.95	41	104,603
Portuguese	0.96	0.90	0.95	0.98	25	5,000
Romanian	0.83	0.97	0.97	0.97	28	a
Russian	0.92	0.88	0.91	0.95	44	188,000
Sanskrit	0.96	0.82	0.82	0.96	48	10,000
Slovak	0.89	0.92	0.95	0.96	44	20,000
Spanish	0.91	0.97	0.97	0.98	24	500,000
Swedish	0.82	0.99	0.98	0.99	41	22,000
Tadzhik	0.92	0.83	0.88	0.93	32	119,648
Ukranian	0.89	0.94	0.97	0.97	39	50,000
Vedic	0.93	0.93	0.93	0.98	43	12,170

Yukaghir-Uralic						
Finnish	0.83	1.00	0.95	1.00	56	73,289
Hanty (Eastern)	0.92	0.87	0.89	0.94	37	110,990
Hanty (Kazym)	0.90	0.94	0.94	0.97	27	74,762
Hungarian (Written)	0.86	0.98	0.97	0.98	59	551,828
Hungarian (Oral)	0.89	0.97	0.96	0.99	57	79,395
Karelian (Tihyan)	0.89	0.97	0.95	0.98	72	217,932
Karelian (Udikov)	0.83	0.99	0.98	0.99	41	62,360
Komi (Zyrian)	0.84	0.97	0.98	0.97	36	80,168
Lopari (Saam)	0.89	0.95	0.92	0.97	67	109,894
Mansi (Sosva)	0.88	0.93	0.96	0.95	28	276,284
Mansi (Konda)	0.94	0.94	0.96	0.99	36	19,287
Mari (Lawn)	0.78	0.98	0.96	0.99	36	105,959
Mari (Mountain)	0.77	0.97	0.94	0.98	34	101,927
Nenets	0.92	0.88	0.93	0.95	35	13,745
Nganasan	0.90	0.90	0.95	0.95	41	a
Selkup	0.94	0.94	0.92	0.99	73	10,000
Udmurt	0.80	0.98	0.96	0.98	36	110,245
Vepsian	0.93	0.95	0.96	0.99	57	153,675
Yukaghir	0.93	0.87	0.90	0.95	39	34,934
Miscellaneous Families						
Arabic	0.90	0.94	0.98	0.97	43	23,727
Burmese	0.94	0.92	0.93	0.99	68	94,972
Chinese (Patumhua)	0.86	0.98	0.98	0.98	40	47,837
Chukot	0.94	0.97	0.99	0.99	20	122,154
Dajak	0.96	0.80	0.85	0.96	21	19,999
Eskimo (Imaklin)	0.92	0.95	0.97	0.97	28	61,964
Georgian	0.96	0.92	0.94	0.98	33	100,000
Hausa	0.97	0.80	0.78	0.97	28	79,790
Hawaiian	0.93	0.87	0.90	0.94	13	a
Hebrew	0.96	0.84	0.90	0.97	26	66,000
Indonesian	0.88	0.83	0.90	0.90	29	72,509
Itelmen	0.89	0.98	0.99	0.99	33	75,198
Kaiwa	0.96	0.92	0.96	0.98	21	a
Ket	0.89	0.95	0.98	0.97	31	33,120
Koryak	0.92	0.92	0.92	0.95	23	147,179
Mangaryi	0.93	0.72	0.76	0.93	22	19,089
Naukan	0.90	0.97	0.98	0.99	34	48,422
Nivh	0.87	0.97	0.98	0.98	38	86,516

Samoan	0.93	0.83	0.87	0.93	15	a
Vietnamese	0.96	0.91	0.85	0.99	74	7,780
Total Mean	0.90	0.93	0.95	0.97	38.8	

a. Not given in source.

Table 1 *Values of R^2 for fit of Zipf, Sigurd, Borodovsky and Gusein-Zade, and Yule equations to phoneme frequencies in 95 languages; and information on number of phonemes and number of phonemes in sample for each language*

A glance at Table 1 shows that the Yule equation generally fits the data best. In order to compare how well the equations describe the phoneme distributions, we used the Sign Test (Siegel & Castellan 1988) to compare R^2 values. This is not an especially powerful test but has the virtue of requiring us to make no assumptions about the distribution of the R^2 values. The Sign Test requires only that pairs of scores being compared be drawn from samples that are comparable in regard to any extraneous factors. This is usually accomplished by comparing scores from the same person (compare language). In any event, we do not need a powerful test to demonstrate the obvious. When the Sign Test is used, tied scores are discarded. Most of what appear to be ties in Table 1 were resolved by examining values of R^2 carried to four decimal places. This is legitimate given that large numbers of phonemes were used in the tabulations for most languages. Usually, ties were broken by examining R^2 carried to only three decimal places. For three comparisons, R^2 values were identical when R^2 was carried to four decimal places. These were regarded as ties and the cases were not included in the analysis. The results are shown in Table 2. As may be seen, the Yule equation is statistically better in all cases as well as for all the language families. The Borodovsky and Gusein-Zade and Sigurd equations are often not significantly different. The Zipf equation is actually better than the Sigurd equation for one language group.

Comparison	Language Group				
	Altaic	Indo-European	Yukaghir Uralic	Miscellaneous	Total
Yule > Zipf	.001	.001	.001	.001	.00001
Yule > Sigurd	.001	.001	.001	.001	.00001
Yule > Borodovsky	.05	.001	.001	.05	.00001
Borodovsky > Zipf	.001	.05	.01	*ns*	.001
Borodovsky > Sigurd	*ns*	.05	*ns*	.05	.00001
Sigurd > Zipf	.01	.05	*ns*	*ns*	.001

Table 2 *Two-tailed probabilities based upon the Sign Test that R^2 values differ for the equations under consideration for the language groups*

The more phonemes there are in a language, the better their distribution is fit by the Sigurd equation, $r(93) = .23$, $p < ,05$, This is also the case for the Yule equation, $r(93) = ,37$, $p < .001$. This is presumably the case because the more phonemes a language has, the more phonemes of low frequency of occurrence it has. At least this is the case for the 95 languages in our sample. These equations are able to capture large variations in frequency better than the others considered. For now, we seem safe in concluding that phoneme frequencies follow a Yule distribution. It fits the distribution of phonemes in the four language groups shown in Table 1 equally well, $F(3,91) = 1,55$, *ns*. In contrast, the other equations fit one group slightly ($p < .05$) better than the others. The Borodovsky and Sigurd equations are best at describing the frequency distributions of the Altaic languages and worst at describing the distributions of the miscellaneous group. The Zipf equation performs best with the miscellaneous group and worst with the Uralic languages. As shown in Table 2, though, the Yule equation provides the best fit for all four language groups. A better equation may be found in the future. We have examined plots of the fit of the equations to the data for all 95 languages. When the Yule equation does not fit the data extremely well, we see no consistent reason why. Deviations between predicted and observed values seem to occur at random places. It does have a tendency to over-predict the values of very low frequencies, as do the other equations, but this is not generally the main reason in cases where it does not provide an extremely good fit to observed values.

It is interesting that the same equation describes the frequency distributions of DNA codons and phonemes. This may be purely a coincidence. However, it might imply a similarity between linguistic and genetic information transmission. Simon (1955) explains Zipf and Yule distributions as likely to arise when the probability of the next symbol in a message is proportional to how often each symbol has previously been used. We could assume a very unlikely stream of speech in which each of the phonemes has been used once. The next phoneme would be chosen at random with each having a probability of $1/n$, where n is the number of available phonemes. The next phoneme would be chosen at random from the set of phonemes, one of which now had a probability of $2/n$ of being chosen. After a few recursions, some phonemes or codons will be very common and some very uncommon. There is a countervailing force for accurate transmission of information that prevents phonemes or codons from becoming too frequent. Phonemes may be distributed in a less skewed fashion than are words because the extreme repetition called for by the Zipf distribution would cause difficulties in comprehension: Frequent repetitions of a phoneme cause people to misperceive it as a similar phoneme (Eimas and Corbit 1973). On the genetic level, the degree of repetition that would be found if a Zipf distribution held for codons would lead to a maladaptive overproduction of some amino acids and underproduction of others. The Zipf distribution works for word frequencies or size of cities because there are a lot of words and a lot of cities. It seems not to be found when there are fewer entities, whether these are phonemes or codons or poets, from which to choose.

References

ALTMANN, G. (1993). Phoneme counts: Marginal remarks on the Pääkkönen article. In ALTMANN, G. (Ed*.), Glttometrika*, *14*. Trier: Wissenschaftlicher Verlag Trier, 54-68.

BORODOVSKY, M., Yu., & GUSEIN-ZADE, S. M. (1989), A general rule for ranged series of codon frequencies in different genomes. *Journal of Biomolecular Structure and Dynamics, 6*, 1001-1013.

DEWEY, G. (1923), *Relative frequency of English speech sounds*. Cambridge, MA: Harvard University Press.

EIMAS, P. D., & Corbit, J. D. (1973), Selective adaptation of linguistic feature detectors. *Cognitive Psychology. 4*, 99-109.

GOOD, I. J. (1969), Statistics of language. In Meethoun, A. R., & Hudson, R. A. (eds.), *Encyclopaedia of information, linguistics and xontrol*. Oxford: Pergamon, 567-581.

GUSEIN-ZADE, S. M. (1987), On the frequency of meeting of key words and on other ranked series. *Science-Technical Information, Series 2: Information Processes and Systems, 1*, 28-32. (in Russian)

GUSEIN-ZADE, S. M. (1988), On the distribution of letters of the Russian language by frequencies. *Problems of the Transmission of Information, 23*, 102-107. (in Russian)

KUČERA, H., & Monroe, G. K. (1968*), A comparative quantitative phonology of Russian, Czech, and German*. New York: American Elsevier.

LOTKA, A. J. (1926), The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*, 317-323.

MARTINDALE, C. (1995), Fame more fickle than fortune: On the distribution of literary eminence. *Poetics, 23*, 219-234.

MARTINDALE, C., GUSEIN-ZADE, S., MCKENZIE, D. P., & BORODOVSKY, M. Y. (1996), Comparison of equations describing the ranked frequency distributions of phonemes and graphemes

Journal of Quantitative Linguistics. 3, 106-112.

MARTINDALE, C., & KONOPKA, A. K. (1996), DNA oligonucleotide frequencies follow a Yule distribution. *Computers and Chemistry, 20,* 35-38.

McNEMAR, Q. (1962), *Psychological statistics*. New York: Wiley.

PARETO, V. (1896), *Cours d'économie politique*. Geneva: Draz.

RUHLEN, M. (1987), *A guide to the world's languages, Vol. 1: Classification*. Stanford: Stanford University Press.

SIEGEL, S., & CAPTELLAN, N. J. (1988*), Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

SIGURD, B. (1968), Rank-frequency distributions for phonemes. *Phonetica, 18,* 1-15.

SIMON, H. A. (1955), On a class of skew distribution functions. *Biometrika, 42*, 425-440.

TAMBOVTSEV, Yuri. (1994a), Dynamics of Functioning of Phonemes in the Speech Chains of the Languages of Different Structure. (in Russian). Novosibirsk: Novosibirsk University Press.

TAMBOVTSEV, Yuri. (1994b), Typology of Orderliness of Speech Chains in Language (in Russian). Novosibirsk: Novosibirsk University Press.

TAMBOVTSEV, Yuri. (2001a), Compendium of the Basic Characteristics of Functioning of Consonants in the Speech Chain of English, German, French and Other Indo-European Languages. (in Russian). Novosibirsk: NCI.

TAMBOVTSEV, Yuri. (2001b). Functioning of Consonants in the Speech Chain of the Ural-Altaic Languages. Novosibirsk: NCI.

TAMBOVTSEV, Yuri. (2001c). Some Theoretical Foundations of the Typology of the Orderliness of Phonemes in the Speech Chain and Compendium of the Statistical Characteristics of the Basic Groups of Consonants. (in Russian). Novosibirsk: NCI.

TAMBOVTSEV, Yuri. (2003a). Typology of Funtioning of Phonemes in the Speech Chain of Indo-European, Paleo-Asiatic, Ural-Altaic and Other World Languages: Compactness of Sub-Groups, Groups, Families and Other Taxa. Novosibirsk: SNI.

TAMBOVTSEV, Yuri. (2003b). Fenno-Ugristica # 25. Measuring Phono-statistical Distances between Uralic Languages. (in Russian). Tartu: University of Tartu Press, pp. 120 – 168.

WEISS, M. (1961), Über die relative Häufigkeit der Phoneme des Schwedischen. *Statistical Methods in Linguistics, 1*, 41.55.

WITTEN, I. H., & Bell, T. C. (1999), Source models for natural language text. *International Journal of Man-Machine Studies, 32*, 545-579.

YULE, G. U. (1924), A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London Biological Sciences, 213,* 21-87.

ZIPF, G. K. (1929), Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology, 40*, 1-95.

ZIPF, G. K. (1949*), Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

Yuri Tambovtsev
Department of English and Linquistics
Novosibirsk Pedagogical University
Russia4111
yutamb@mail.ru

Colin Martindale
Department of Psychology
University of Maine
North Drinkwater Blvd., Apt. B406
Scottsdale AZ 85251
U.S.A.
cmartin61@cox.net

In *SKASE Journal of Theoretical Linguistics* [online]. 2007, vol. 4, no. 2 [cit. 2007-06-14]. Available on web page <http://www.skase.sk/Volumes/JTL09/pdf_doc/1.pdf>. ISSN 1336-782X.