Glottometrics 22
2011

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 22 (2011), Lüdenscheid: RAM-Verlag, 2011. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 22 (2011) **ISSN 2625-8226**

Contents

Glottometrics 22, 2011, 1-4

Diversification of a single sign of the Danube script

Karl-Heinz Best

Abstract. In this paper the positive binomial distribution has been fitted to the ranked distribution of variations of a character of the Danube script. It brings a further corroboration of the hypothesis that diversification processes abide by laws.

Keywords: Danube script, laws, diversification, binomial distribution

1. The Danube script

In the present contribution we study the diversification of a single sign in the Danube script. The Danube script is a sign system whose interpretation is not unequivocal. The signs were found in South East Europe and can be dated approximately to 5300 to 3200 B.C. (Haarmann 2010: 10). In the literature, they are known under different names such as "old European script" (Dürscheid 2006: 104), "old Balkan script" (Haarmann 1990: 73), or the script of the "Vincha culture" (Haarmann 1990: 73) because of its place of discovery in the vicinity of Beograd.

According to Haarmann these signs represent a script. He claims that both positive and negative criteria testify to the fact that the signs may represent a script; and that the negative evidence, viz. the lack of symmetry, and the positive one, viz. the repetition of signs, indicate that the signs are not ornaments or motives with decorative function (Haarmann 2010: 29). The reason for the script theory is the linear ordering of the signs and the specific way of linking them with artefacts (Haarmann 2010: 31). The opposite view is expressed by Dürscheid (2006: 105) as follows: "Ein Grund hierfür ist sicher darin zu sehen, dass die Antwort auf die Frage nach den Anfängen der Schrift eng damit zusammenhängt, welchen theoretischen Standpunkt man in Bezug auf die Definition von Schrift einnimmt."

The Danube script – if it is really a script – is a logographic one, not a letter script; it does not reflect the sound form of the words. This fact renders the interpretation of the script difficult. So far one can distinguish more than 700 signs (Haarmann 2010: 23), some of which are pictorial (iconic), so that at least with a part of these signs an interpretation seems to be possible.

2. The distribution of variants of a selected sign of the Danube script

We shall not continue pursuing the question whether it is a script or not. The problem tackled here does not depend on this decision. Haarmann (2010: 68) collects in a table the signs whose simplest form is a capital "V" ("basic sign") and the other ones differ from it by additional strokes as simple or complex variations. This sign is interpreted by some researchers as the logogram of divinity.

If one considers the total of 29 signs which represent the basic sign or its variations, evaluates them following the proposal by Köhler (2008: 6f.) and orders them according to the number of strokes of each of the signs, one obtains a table with totally 15 ranks. Yu (2001)

considering the different number of strokes in signs of Chinese texts has shown that this method yields good results. (Japanese: Sanada 2008, 99-102). The basic sign "V" has two strokes and has the rank 1, signs with one additional stroke obtain the rank 2, those with two additional strokes rank 3, etc. Both the placing and the length of the additional strokes are neglected. Only in one case the additional stroke is an arc, all other ones are straight lines. Hence the criterion of sign complexity is merely the change of direction of the line. We have to do here with the diversification of signs according to their complexity, a problem proposed for solution by Strauss, Fan & Altmann (2008: 9).

Diversification is a law-like process stated several times by Altmann (1985, 1991) and substantiated in many investigations (cf. references in Altmann 2005). In order to model this phenomenon, one can choose different probability distributions (Altmann 2005: 649ff.) but one may model it also by a continuous function. The causes of diversification are not known, hence it is not possible to select a distribution reflecting the given boundary conditions. One must take into account several possibilities and proceed inductively, an easy way made possible by the software Altmann-Fitter (1997). In the present case it can be shown that the positive binomial distribution defined as

$$P_x = \binom{n}{x} \frac{p^x q^{n-x}}{1-p^n}, \quad x = 1,2,...,n$$

is an adequate way to capture the variations of the sign "V". The fitting of this distribution yielded the result presented in Table 1.

Table 1
Fitting the positive binomial distribution to the variations of the sign "V" in the Danube script (Haarmann 2010: 68)

Rank	n_x	NP_x	Rank	n_x	NP_x
1	1	2.05	9	0	0.03
2	5	4.94	10	1	0.00
3	7	7.16	11	1	0.00
4	6	6.92	12	0	0.00
5	5	4.68	13	0	0.00
6	0	2.26	14	0	0.00
7	1	0.78	15	1	0.00
8	1	0.19			
n = 11, p = 0.3256, X^2 = 1.616, df = 3, P = 0.66					

Legend:

Rank: variation of the sign "V" according to the number of strokes of which the sign consists, beginning with the simple sign "V" having two strokes with rank 1; rank 2: sign "V" with three strokes, etc. The ranks represent the extent of complexity of the variant of "V". As a matter of fact, *rank = complexity* – 1.

n, p: parameters of the distribution

n_x: the frequency of the given variant of "V"

NP_x: computed frequency of signs with the given complexity

df: degrees of freedom (several classes have been pooled, hence there are only 3 df)

X^2: chi-square value

P: the exceedance probability of X^2

We consider the fitting as acceptable if $P \geq 0.05$. Since this condition is fulfilled we may accept the binomial distribution as a model of complexities of the "V"-sign. A graphical presentation can be seen in Figure 1.

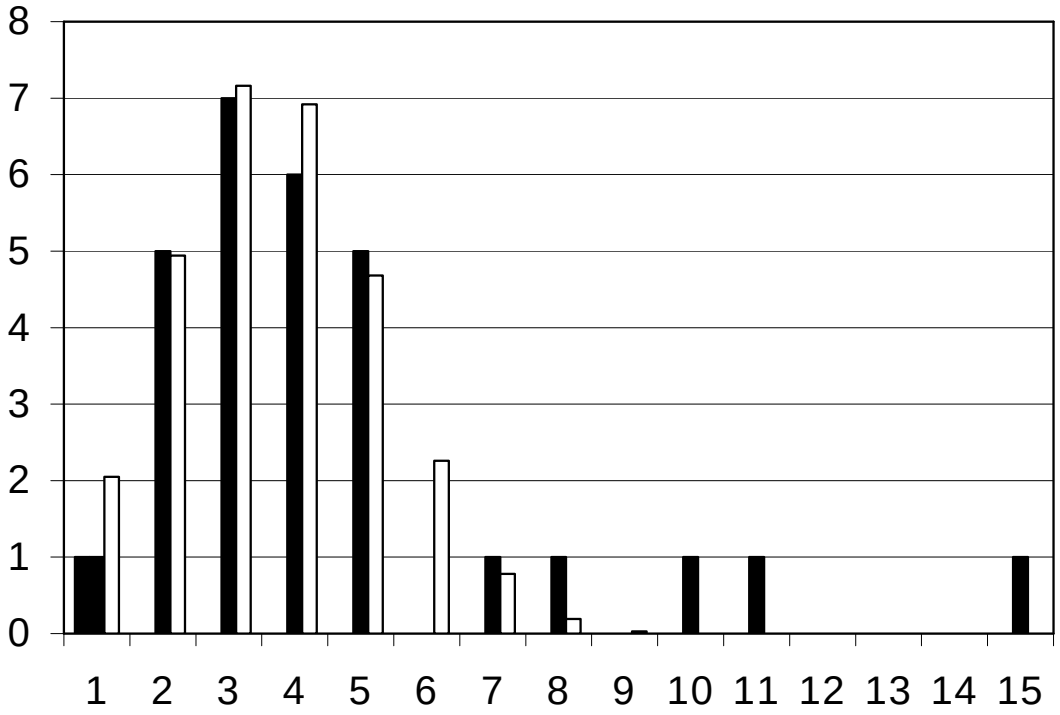

Figure 1: Fitting the positive binomial distribution to the variations of the sign "V"

3. Conclusions

It can be stated that the inventory of variants of "V" ranked according to their complexity abide by the positive binomial distribution, one of the diversification models. This does not depend on the decision whether these signs are really a script or something else. The only important fact is that the signs follow one of the forms of the diversification law which seems to hold even if applied to a phenomenon whose nature is still in controversy.

If some time in the future the consensus in the field is that the Danube script is not a script but an artistic product, this would not change anything in its distribution. Since Orlov et al. (1982) it is known that there are laws even in arts.

References

Altmann, Gabriel (1985). Semantische Diversifikation. *Folia Linguistica XIX, 177-200.*

Altmann, Gabriel (1991). Modelling diversification phenomena in language. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar: 33-46.* Hagen: Margit Rottmann Medienverlag.

Altmann, Gabriel (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Hrsg.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch: 646-658.* Berlin/ NewYork: de Gruyter.

Dürscheid, Christa (2006). *Einführung in die Schriftlinguistik.* 3., überarbeitete und ergänzte Auflage. Göttingen: Vandenhoeck & Ruprecht.

Haarmann, Harald (1990). *Universalgeschichte der Schrift.* Frankfurt/New York: Campus.

Haarmann, Harald (2010). *Einführung in die Donauschrift.* Hamburg: Buske.

Köhler, Reinhard (2008). Quantitative analysis of writing systems: an introduction. In: Altmann, G., Fan, F. (eds.), *Analyses of Script. Properties of Characters and Writing Systems: 3-9.* Berlin/New York: Mouton de Gruyter.

Orlov, Jurij Konstantinovič, Nadarejšvili, Isabella Šotaevna, & Boroda, Mojsej Grigor'evič (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Strauss, Udo, Fan, Fengxiang, & Altmann, Gabriel (2008). *Problems in Quantitative Linguistics I.* Lüdenscheid: RAM-Verlag.

Sanada, Haruko (2008). *Investigations in Japanese Historical Lexicology (Revised Edition)* (S. 97-99). Göttingen: Peust & Gutschmidt.

Yu, Xiaoli (2001). Zur Komplexität chinesischer Schriftzeichen. *Göttinger Beiträge zur Sprachwissenschaft 5, 121-129.*

Software

Altmann-Fitter. 1997. *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Glottometrics 22, 2011, 5-16

Euphony in Slovak lyric poetry

Radek Čech
Ioan-Iovitz Popescu
Gabriel Altmann

1. Total euphony

Euphony is a well known phenomenon in poetry, especially since the introduction of rhyme which directly evokes it. In classical Javanese poetry either the vowels had their prescribed place in the line or they expressed some special mood. In poetic constructions of this kind the placing of some sounds is conscious and represents some kind of binding. However, there are also cases where the poet is not quite aware of the phonic component of his text; he cares for the content, but subconsciously he creates a construction containing elements of euphony. If we want to demonstrate it, we must set up a definition of euphony and present a method of its measurement.

Euphony in a line will be defined here as a function of non-random (i.e. significant) repetition of one or more sounds. Every phoneme can contribute to euphony either by its special position in the line or by its mere repetition. Some sounds may contribute to the euphony of the strophe (e.g. those in the rhyme position) but need not have a euphonic value in the line. The same holds for a combination of two or more sounds in a certain order.

In general, a sound can have a euphonic value only if it occurs in the line at least twice. Since neither vowels nor consonants can alone fill the whole line, we shall consider separately the number of vowels (*V*) and that of consonants (*C*). If there are *V* places for vowels, then a given vowel has 2^V possibilities of appearing there in different combinations (positions). But this fact must be scrutinized for each vowel separately because for all of them the probability of their occurrence is different. A sound occuring seldom in the language has a smaller probability, but the smaller the probability, the greater is the euphonic values of its appearance. We simplify the problem and consider the occurrence at any position as independent from previous occurrences. Then a sound can occur in the given position with probability p and fail to occur with probability $1 - p = q$. Hence the probability that a vowel occupies x positions out of V possible ones is given as

$$(1) \qquad P_x = \binom{V}{x} p^x q^{V-x},$$

i.e. binomially. The same holds for consonants replacing *V* by *C*. Since we need not only the probability of the given occurrence but also all the extreme ones, we compute (cf. Altmann 1966, Wimmer G. et al. 2003; Strauss, Fan, Altmann 2008: 45f)

$$(2) \qquad P(X \geq x_i) = \sum_{x=x_i}^{V} \binom{V}{x} p^x q^{V-x} ,$$

where x_i is the empirically observed number of the vowel *i* in the line (replace *V* by *C* with consonants). Setting the significance level at $\alpha = 0.05$ we consider a sound as having a euphonic value only if $P(X \geq x) < 0.05$. In order to express the euphonic weight of the given probability, we set up the indicator

$$(3) \qquad E(i) = \begin{cases} 100[\alpha - P(X \geq x_i)], & if \ \alpha > P(X \geq x_i) \\ 0 \quad otherwise \end{cases},$$

and compute it for each sound of the line. In order to obtain the euphonic value of the line, we compute the mean euphony $E(i)$ of those phonemes which have a positive euphony (> 0). Hence for a line we obtain

$$(4) \qquad \bar{E}(line) = \begin{cases} \frac{1}{k} \sum_{i=1}^{k} E_i, & if \ k \geq 1 \\ 0 & otherwise \end{cases}$$

where *k* is the number of sounds with significant euphony (not all sounds of the line). This indicator can be used for studying the course of euphony in the poem.

Let the number of verses in the poem be *n*. Then the extent of euphony in the poem can be obtained from the formula

$$(5) \qquad E(poem) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{k} \sum_{i=1}^{k} E_{ij} = \frac{1}{n} \sum_{j=1}^{n} \bar{E}(line)_j$$

which simply adds the euphonic results in all lines and divides the sum by the number of lines.

Obtaining the expected relative frequencies *p* of individual sounds is a problem of its own. If one analyzes the work of a poet who lived in 19th century, one cannot use the frequencies obtained from a modern corpus. Since in language there are no populations (cf. Orlov, Boroda, Nadarejšvili 1982); one cannot improve the facts taking a corpus of texts from the 19th century. For some centuries (or languages) there were even no corpora. One can approximate the quasi-population by choosing the best available texts, viz. the works of the given author but if (s)he wrote both prosaic and poetic works, mixing them up may strongly change the (expected) proportion. Hence the best approximation to the expected relative frequencies is given by the works of the same author written in the same text sort. However, since we consider vowels and consonants separately, we must use the conditional probabilities, i.e. $p_{vowel} = f_{vowel}/N_V$.

Our aim is to study the euphony in the poetic work of the Slovak poetess Eva Bachletová. To this end we computed the frequencies of sounds in all her poems we

want to analyze and set up a table of observed frequencies from which we obtained the expected relative frequencies as f_x/N where N is either N_V or N_C for vowels and consonants respectively. However, we first reorganized the Slovak sound inventory from the euphonic point of view which differs from the inventory presented in http://www.ui.sav.sk/speech/sampa_sk.htm (retrieved April 5, 2011).

Since short and long vowels do not differ acoustically, we consider [a] and [a:] etc. euphonically as identical and obtain the vowels [a,e,i,o,u]; the vowel written as {ä} is pronounced as [e] both in the language of the poetess (personal communication) and in the mother tongue of one of the authors. There are four diphthongs [ia, ie, iu, uo]; the diphthong written as {ov, ou} is interpreted by two sounds [o] and [v]. Consonants {l,ĺ,r,ŕ} and their syllabic forms are unified and represented by the short forms [l, r]. The three variants of {n}, i.e. {n,N\,N} are unified in [n]. The final {-j}, in SAMPA written as {i_^} is identified with {j}. The SAMPA {G} is identical with {ch}, phonetically [x], and the SAMPA {F} has been eliminated because it represents the nasalisation of [a]. Thus we obtain the sounds and their frequencies as presented in Table 1. It is to be emphasized that we do not perform phonemic but rather euphonic analysis. The sample consists of 8515 sounds.

Table 1
Frequencies of sounds in E. Bachletová's poems

Sound	Freq.	Rel. freq.	Rel.freq. category	Sound	Freq.	Rel. frequency	Rel.freq. category
a	913	0,10722255	0,25396384	n	307	0,03605402	0,06239837
e	792	0,09301233	0,22030598	s	444	0,05214328	0,09024390
o	777	0,09125073	0,21613352	z	167	0,01961245	0,03394309
i	673	0,07903699	0,18720445	l	154	0,01808573	0,03130081
u	270	0,03170875	0,07510431	r	375	0,04403993	0,07621951
ia	38	0,00446271	0,01057024	c	193	0,02266588	0,03922764
ie	116	0,01362302	0,03226704	J\ = ď	74	0,00869055	0,01504065
iu	1	0,00011744	0,00027816	S = š	93	0,0109219	0,01890244
uo	15	0,0017616	0,00417246	Z = ž	93	0,0109219	0,01890244
p	252	0,02959483	0,05121951	tS = č	90	0,01056958	0,01829268
b	182	0,02137405	0,03699187	dZ=dž	8	0,00093952	0,00162602
f	73	0,00857311	0,0148374	L = ľ	141	0,01655901	0,02865854
v	283	0,03323547	0,05752033	j	197	0,02313564	0,04004065
w	54	0,00634175	0,01097561	J = ň	205	0,02407516	0,04166667
m	374	0,04392249	0,07601626	k	262	0,03076923	0,05325203
t	355	0,04169113	0,07215447	g	24	0,00281856	0,00487805
d	212	0,02489724	0,04308943	h	133	0,0156195	0,02703252
ts	85	0,00998238	0,01727642	x	83	0,0097475	0,01686992
dz	7	0,00082208	0,00142276				

Rel.freq. category = conditional relative frequency within the category V or C.

For the sake of illustration we show the analysis of the first line of the poem *Aby spriesvitnela.* Orthographically we have

Nemám rada bielu

(eu)phonically.we may write

[ňemam rada bielu].

Since only sounds occurring twice or more can be taken into account, we have two candidates: [m] and [a]. There are 7 consonants in the line, hence $C = 7$; the conditional probability of [m] is 0,07215447 as shown in Table 1, hence we compute

$$P([m] \geq 2) = \sum_{i=2}^{7} \binom{7}{i} 0,07215447^{\,i} (1 - 0,07215447)^{7-i} = 0,09389934.$$

Since the result is greater than 0,05, the euphonic weight is zero.

Further, we have [a] three times, $V = 6$, hence

$$P([a] \geq 3) = \sum_{i=3}^{6} \binom{6}{i} 0,25396384^{i} (1 - 0,25396384)^{6-i} = 0,17575387$$

again, $E(a) = 0$. In the third line of the poem we find the first significant frequency, namely with [ň]. There are 7 consonants and twice [ň], hence

$$P([\check{n}] \geq 2) = \sum_{i=2}^{7} \binom{7}{i} 0,04166667^{i} (1 - 0,04166667)^{7-i} = 0,0317008$$

yielding

E([ň]) = 100(0,05 – 0,0317008) = 1,8299.

The results are presented in Table 2 together with the orthographical text of the poem.

For easier checking, the number of consonants (C) and vowels (V) in the line is given, too.

Table 2
Euphonic values of sounds in the poem *Aby spriesvitnela*

Text	**C**	**V**	**Euphonies**
Nemám rada bielu	7	6	
dnes je prízrakom chladu	13	7	
znecitlivenia	7	5	[ň] = 1,8299
konečného verdiktu	10	7	
nad človekom	7	4	
nad pocitom	6	4	
nad láskou.	6	3	
Dnes je tu iná biela	8	6	
biela obrazovky	7	6	[b] = 2,4616
počítača	4	4	[č] = 4,8041
tam nahadzujeme	7	6	
svoje vnemy	6	4	
čiernymi linkami	8	6	[i] = 3,6665
rýchlo a bezpečne	8	6	
kreslíme životy	8	6	
slovami,	4	3	
ktoré navždy	7	4	
zmenili bielu	6	5	
a odviedli nás	6	5	
od základných farieb	10	6	
bytia.	2	2	
A možno stačí jedna	9	7	
nenapísaná veta	7	7	
aby „novodobá“	5	6	[b] = 3,7301
biela spriesvitnela.	10	6	[l] = 1,2696, [ie] = 3,5678
Lebo čistá – biela krehkosť	13	8	
prichádza potichu...	7	6	[p] = 0,3620, [x] = 4,4351

The above table yields a number of research possibilities: (1) Is there some regularity in the values of euphony in the course of the poem? (2) What part of euphonies is made up by the sounds of the poem title? (3) How to compare statistically the euphony of two poems? (4) Is there some historical evolution of euphony in the work of the given author?

Here we shall consider only the overall euphony of the poem. Using formula (5) we add all mean euphonies of lines and divide by the number of lines (n = 27). We obtain

$$\bar{E}(poem) = [1{,}8299 + 2{,}4616 + 4{,}8041 + 3{,}6665 + 3{,}7301 + (1{,}2696 + 3{,}5678)/2 + (0{,}3620 + 4{,}4351)/2]/27 = 21{,}3095/27 = 0{,}7892,$$

which represents the mean euphony of the poem per line. As can be seen in Table 2, many lines have E(line) = 0. The euphonies occurring in the same line are averaged, not added.

The variance of the poem's euphony which will be used in comparisons can be computed in different ways according to the aspect of comparison. We restrict ourselves to the following procedure. We compute the mean squared deviations of mean line euphonies from $\bar{E}$ (*poem*), i.e.

$$(6) \qquad Var(E) = \frac{1}{n-1}\sum_{i=1}^{n}[\bar{E}(line)_i - \bar{E}(poem)]^2$$

yielding in the above case

$$Var(E) = [20(0\text{-}0{,}7892)^2 + (1{,}8299 - 0{,}7892)^2 + (2{,}4616 - 0{,}7892)^2 + (4{,}8041 - 0{,}7892)^2 + (3{,}6665 - 0{,}7892)^2 + (3{,}7301 - 0{,}7892)^2 + (2{,}4187 - 0{,}7892)^2 + (2{,}3986 - 0{,}7892)^2]/26 = 2{,}1011,$$

and this variance can be used in the asymptotic normal test for comparing the mean euphonies of two poems, namely as

$$(7) \qquad u = \frac{|\bar{E}(poem)_1 - \bar{E}(poem)_2|}{\sqrt{\frac{Var(E_1)}{n_1} + \frac{Var(E_2)}{n_2}}}.$$

The division of $Var(E)$ by the number of lines must be performed because we need $Var(\bar{E}(poem))$. For example, the difference between the poems *Aby spriesvitnela*.and *Iba neha* yields

$$u = \frac{|0.7894 - 0.8698|}{\sqrt{\frac{2.1011}{27} + \frac{2.5877}{54}}} = 0.2267$$

which is not significant

Table 3
Values of euphony in poems by E. Bachletová

Poem	**#Lines**	**#Euphonies**	$\bar{E}(poem)$	***Var*(*E*)**
Aby spriesvitnela	27	9	0,7892	2,1011
Bez rozlúčky	16	6	0,7316	1,8172
Čakáme šťastie	13	9	1,1597	1,3579
Čakanie na boží jas	29	5	0,4194	1,3195
Čas pre nádych vône	18	10	1,2001	2,6953
Dielo Stvoriteľa	44	19	0,7815	2,0019
Dnešný luxus	12	5	1,3363	3,4899
Do večnosti beží čas	18	5	0,5501	1,3292
Ešte raz	7	5	2,2522	3,2378
Hľadanie odpovedí	24	11	0,9947	2,1287
Iba neha	54	19	0,7784	2,2743
Iba v modlitbe	5	7	1,8158	1,3932
Iba život	14	29	2,6472	1,5591
Ihly na nebi	21	7	1,8567	1,0482
Istota	9	2	0,6610	1,9158
Každodennosť	8	6	1,9020	4,3369
Keď dohorí deň	14	10	1,6372	2,0889
Kým ich máme	16	3	0,4581	1,9220
Malé modlitby	11	21	2,4515	3,4998
Mladé oči	7	2	0,6855	1,4395
Malý ošiaľ	27	12	0,8494	2,1483
Moje určenie	52	17	0,4907	1,1333
Nado mnou ty sám	10	4	0,9285	2,7892
Náš chrám	23	13	1,1775	2,9273
Naše dejiny	7	5	1,0391	2,0358
Naše mamy	14	4	0,9628	2,8418
Naše svetlo	28	17	1,5164	3,3362
Návraty	8	4	0,9654	2,7769
Neha domova	9	6	1,4756	4,0364
Neopusť ma	6	5	1,8638	2,3396
Nepoznateľné	51	19	0,6472	1,5851
Otázka	6	5	1,1409	2,1372
Podobnosť bytia	12	3	0,2917	0,3513
Precitnutie	13	6	0,8343	1,9732
Prvotný sen	27	15	1,1919	2,4199
Rozdelená bytosť	26	8	0,4548	1,0741
Rozťatá prítomnosť	36	8	0,5446	1,7323
Smútok	9	3	0,6482	1,2754
Som iná	21	5	0,4125	1,2397
Spájania	14	4	0,1713	0,2392
Stály smútok pre šesť písmen	48	15	0,4819	0,9211

Tak málo úsmevu	20	11	1,1373	2,5246
Tiché verše	12	2	0,3162	1,0939
To všetko je dar	24	8	0,6882	1,9045
Večerná ruža	15	10	1,4516	3,7202
Večerné ticho	19	39	2,6204	1,2604
Vo večnosti slobodná	31	9	0,8760	2,6253
Vrátili sa	12	4	0,5027	1,1785
Vyznania	26	5	0,3935	1,3004
Z neba do neba	40	13	0,6763	1,7015
Zasľúbenie jasu	12	9	1,8127	2,5282
Zázrak	6	1	0,5695	1,9458
Zbytočné srdce	11	5	0,7625	1,3176

Looking at the numbers in Table 3 we can state that euphony is a quite irregular phenomenon. It is created ad hoc, sometimes subconsciously, sometimes consciously and in many cases in dependence on the meanings of words which are more important than their phonic structuring. In rhymed poetry it has a greater importance because, at least subconsciously, the poet searches for phonic agreements and tries to place the pertinent words at the end of the lines.

It can be shown that the extent of euphony, $\bar{E}(poem)$, does not depend on the number of lines in the poem – at least not for Bachletová – and the same holds for the number of euphonic expressions in relation to the number of lines. The line euphonies of Bachletová lie in the interval <0,3162; 2,6472>, i.e., rather in the lower part of euphony which may move in the interval <0; 5>. This state is most probably caused by the small length of her verses containing many times only one word. According to personal communication, she writes her poems "in one go" and makes corrections only in exceptional cases.

If we observe the phones and the sums of their euphonic values in all poems as presented in Table 4, we can state that no phonetic/phonemic order can be discerned. The most striking is the fact that within the line, vowels do not play any special euphonic role. There is no preference for voiced or voiceless consonants or for a special place of articulation. Some sounds do not even appear in a euphonic role in the complete work of the author, In order to detect the preferences of the poetess, other poets should be analyzed using the same method. Of course, it would be possible to study the iconic origin (cf. Koch 2005) of some words in which euphonic sounds occur; but the data in Table 4 represent the overall euphonic weights occurring in all words of all poems. The poetess avoids iconism, rhythm – which is also an imitation of a natural phenomenon – and even the modern poetic form expressed by rhyme. She does not express her ideas materially but rather philosophically, not caring for any formal restrictions. Of course, a thorough study of the poems from another point of view could reveal iconic components which are present in all languages, but they do not seem to be present in these poems. The individual sounds having euphonic weight do not have submorphemic meaning, which can be found in many English word beginnings (cf. Lvova 2011) and is object of intensive investigation. "Phonetic meaning" seems to be foreign to these poems.

As a matter of fact a more thorough computation of euphony would be possible. Separating vowels and consonants one could obtain e.g. all vowels that occur at least twice (= euphonic vowels) and compute the cumulative probability of the multinomial distribution in which non-euphonic vowels (= occurring only once) would be joined in a common class. The same could be done separately for consonants. However, such a procedure is not only considerably more complex but it would not yield "better" results (cf. Wimmer et al. 2003).

2. Verse alliteration

Alliterations at the beginning of verses (Skinner version) may appear without the conscious will of the poet. They may be caused by perseveration, formal strengthening, spontaneity, or by grammatical necessity. Nevertheless, the poet can create them consciously, too. Since it is not (always) possible to know the cause, our problem is merely to state whether the status quo can be considered random or whether there is some tendency. Since sounds must be repeated, their repetition is necessary but it may be random. The longer a poem the more frequently all sounds occur and alliteration must arise automatically.

In order to test the existence of a tendency we use the same method as above. However, here we shall use the non-conditional frequencies

Consider again the poem *Aby spriesvitnela* in which the initial sounds of verses are as follows

[N,d,z,k,n,n,n,d,b,p,t,s,tS,r,k,s,k,z,a,o,b,a,N,a,b,L,p]

Combining identical sounds together we obtain the frequencies

t	tS	r	L	o		N	d	z	p	s		k	n	b	a									
1	1	1	1	1		2	2	2	2	2		3	3	3	3									

Since the poem has 27 lines, we use formula (2) replacing V by 27 and the individual *p*'s by the relative frequencies in the second column of Table 1. We obtain only one euphonic tendency, namely with [b] yielding *EU*(b) = 3,0523, hence $\bar{E}(poem)$ = 3,0523/27 = 0,1131. That means, in this poem the alliteration of lines is very low.

Performing this investigation for 54 poems we obtain the results presented in Table 4

Table 4
Verse alliterations in poems by E.Bachletová

Poem	**Lines**	**Significantly alliterated sounds**	$\bar{E}$
Aby spriesvitnela	27	b	0,1131
Bez rozlúčky	16	b,ž	0,2239
Čakáme šťastie	13	f,S	0,6617
Čakanie na boží jas	29	a,n,p	0,3825
Čas pre nádych vône	18	-	0
Dielo Stvoriteľa	44	d,j,p,n	0,3261
Dnešný luxus	12	-	0
Do večnosti beží čas	18	u,L,b	0,591
Ešte raz	7	p	0,7024
Hľadanie odpovedí	24	g	0,1996
Iba neha	54	a,t,tS	0,1887
Iba v modlitbe	5	-	-
Iba život	14	z	0,1433
Ihly na nebi	21	n,J	0,3829
Istota	9	u	0,2088
Každodennosť	8	u	0,3152
Keď dohorí oheň	13	p	0,339
Kým ich máme	16	h	0,1543
Malé modlitby	11	J	0,4545
Malý ošiaľ	27	a	0,1078
Mladé oči	7	t	0,2609
Moje určenie	52	a,f,v,k	0,1766
Nado mnou Ty sám	10	p,d	0,4819
Náš chrám	23	a,p,v	0,2088
Naše dejiny	7	ď=J\	0,6923
Naše mamy	14	-	-
Naše svetlo	28	j,k,d	0,4153
Návraty	8	d	0,4287
Neha domova	9	k	0,2276
Neopusť ma	6	J=ň	5,0000
Nepoznateľné	51	n,L	0,1051

Otázka	6	-	0
Podobnosť bytia	29	z	0,1074
Precitnutie	13	b,h	0,4042
Prvotny sen	27	v,tS,z,	0,4444
Rozdelená bytosť	26	Z,J	0,167
Rozťatá prítomnosť	36	g,J(=ň),p,Z	0,482
Smútok	9	-	0
Som iná	21	s	0,2196
Spájanie	14	j	0,3306
Stály smútok pre šesť pismen	48	a,k,f,J(=ň)	0,3978
Tak málo úsmevu	20	s	0,2346
Tiché verše	12	b	0,1987
To všetko je dar	24	p,t,z	0,3774
Večerná ruža	15	j,p,	0,2996
Večerné ticho	19	f	0,2603
Vo večnosti slobodná	31	tS,J(=ň)	0,1372
Vrátili sa	12	f	0,3785
Vyznanie	26	t,Z	0,3585
Z neba do neba	40	d,p,x	0,2393
Zasľúbenie jasu	12	-	0
Zázrak	6	k,s	0,8582
Zbytočné srdce	11	d	0,1876

Aliteration plays a still more irrelevant role in Bachletová poems than the general line euphony. Though here, the alliteration attained both of its extreme values, i.e. 0 in *Naše mamy* and 5,0 in *Neopusť ma*, the rest of the values is very small. The mean alliteration of 64 poems is $\overline{E}$ = 0,3764 with variance of the mean V($\overline{E}$) = 0,0088

3. Conclusions

The results show that euphony is present in poems even if the author does not create it consciously. It awards to poetry a special phonic colour which need not be present in other text sorts. However, extensive investigation is necessary taking into account both different text sorts and languages. Here we can formulate at least some hypotheses to be tested:

(a) Poetic texts are more euphonic than prosaic ones. Here a test for the comparison of texts must be developed.

(b) The more sounds there are in a language, the more clearly euphony can be presented. The reverse hypothesis would not be quite correct because the computations of euphony is based on sound frequencies and the greater the sound inventory, the smaller are the relative frequencies of sounds and the greater is the chance to obtain small cumulative binomial probabilities.

(c) The measurement of euphony depends also on the aspect we choose. Languages having many (or all).words ending with a vowel and having very few consonant clusters have a more melodious sounding (e.g. Italian, Hawaiian) than those full of consonant clusters, e.g. the Czech sentence "Strč prst skrz krk" will surely not evoke euphonic impressions even if [r] occurs in it frequently; but fortunately, there are no such sentences in texts. Thus, euphony is just a concept having many possible definitions and ways of computation.

References

Altmann, G. (1966). The measurement of euphony. In: *Teorie verše I, 259-261*. Brno: Universita J.E. Purkyně.

Knauer, K. (1969). Die Analyse von Feinstrukturen im sprachlichen Kunstwerk. In: Kreutzer, V, Gunzenhäuser, H. (eds.), *Mathematik und Dichtung: 193-210*. München: Nymphenburger.

Koch, W. (2005). *The iconic roots of language. Essayes on the non-arbitrary origins of human communication*. Lüdenscheid: RAM

Lvova, N.L (2011). Determining phonetic symbolism of the text. In: Kelih, E. et al. (eds.), *Issues in Quantitative Linguistics, Vol 2, 94-103.*

Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Skinner, **B.F**. (1939). The alliteration in Shakespeare's sonnets: A study of literary behaviour. *The Psychological Record 3, 186-192.*

Skinner, B.F. (1941). A quantitative estimate of certain types of sound patterning in poetry. *The American Journal of Psychology 54, 64-79.*

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics Vol 1*. Lüdenscheid: RAM.

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov*. Bratislava: Veda.

Glottometrics 22, 2011, 17-43

Academician Andrey Nikolaevich Kolmogorov as a Scholar of Verse Theory

Vadim Baevskii

In memory of Marina Krasnoperova –
the brave truth-seeker in the verse theory

During the 1960s and 1970s A.N. Kolmogorov, one of the greatest mathematicians of the twentieth century, devoted much of his attention to studying verse theory by means of statistics and probability theory. Together with his students he examined different systems of Russian versification, developing the axioms for its study, and also lent support to employing exact methods of research in the field of literary theory.

Andrey Nikolaevich Kolmogorov was born 25 April 1903 and died 20 October 1987. Since that time, as his students say, a new scientific discipline – kolmogorovedenie, or Kolmogorov studies – has been developing in our country.

The people who were close to him, hisstudents , wrote in their memoirs and said when we met that verse study was a favorite occupation of this preeminent mathematician – and one should keep in mind that he had broad interests in extremely diverse fields of mathematics, in many practical applications of mathematical methods, and in a great variety of scientific schools and realms of culture. Thus, G.N. Barenblatt, an outstanding and world-famous mathematician, writes: "I remember how Natalya Grigorievna Khimchenko – or simply Natasha Rychkova, as all the pupils of Andrey Nikolaevich know her, – a very young woman then, after her graduation from the Mechanics and Mathematics Department, first came to A.N. to help him in his favorite pursuit: statistical approaches to the study of verse" [1, p. 96].

During World War II Kolmogorov served in the military and was awarded seven orders of Lenin. His diary of 1943 has been published – it was written at a crucial time, during the turning-point in the war: the retreat of the German army had begun and major battles were taking place. The diary gives roughly equal attention to four topics: mathematics, German poetry, music, and verse theory. "I've been so strangely created – he writes – that formal analyses of rhythms and other things of that kind apparently even helped me to penetrate into the essence of Goethe's poetry. Anyway, I am absolutely carried away by it at the moment" [2, p. 54].

■□

There were occasional rumors that Kolmogorov was writing a monograph to summarize the results of his research in the field of verse theory. Sometimes he used to write about this himself. For example, on the first page of a study (Kolmogorov 1968: 145) he says: "This article is an extract from a large unfinished work that was conceived as a guide for beginning students of versification". To see his fundamental book on verse theory in print, even unfinished, would have been of great help in understanding ourselves and the development of our peculiar field of scholarship which lies at the intersection of literary theory, linguistics and mathematics… But, unfortunately, we have never gotten this book. And though quite a lot of years have passed since Kolmogorov's death, his

students have not yet written any articles summarizing their teacher's papers in his beloved field of scholarship. Indeed, to do so is not an easy task.

■□

A journal once came out with a one hundred page publication, in which Kolmogorov's analysis of poetic speech is presented against the background of his ideas about semiotics and cybernetics (Uspenskii 1997). Several pages of Kolmogorov's texts simply drown in the wordy, uninformative comments. And when reading these comments you suddenly come across a line taken from an elegy by Pushkin, moreover, not an unknown elegy, hidden somewhere amidst unfinished texts in the depths of the poet's complete works, but from a widely anthologized elegy, "I still recall the wondrous moment when you appeared before my sight …" ("Ia pomniu chudnoe mgnovenie: peredo mnoi iavilas' ty…[1]"), and this very line is said to be Blok's (Uspenskii 1997:. 133–134), you can't help asking yourself whether you can trust the rest of the information. And yet, this is the first attempt to characterize Kolmogorov's studies in the field of verse theory.

The sixth issue of the *Bulletin of Moscow University (Philology)* for 2009 contained a fine report regarding a round-table conference on mathematics and philology that was dedicated to celebrating the seventieth birthday of the University rector, V.A. Sadovnichy. Almost all its participants either directly characterize the activity of Kolmogorov as a scholar of versification or somehow touch upon this subject. Due to the general atmosphere of the meeting, the accounts in the journal are of a popular nature and draw in the reader. The following episode demonstrates the importance of such meetings. As an example of a mathematical truth the following statement is considered: the sum of the angles of any triangle is equal to two right angles. As an example of a humanistic or philological truth – both terms are used by the author to denote one and the same notion – the following statement is given: "Mayakovsky has always been and remains the best and the most talented poet of our Soviet epoch." The first statement, by the way, represents a logical sophism. The author explains that this statement is true according to Euclidian geometry but is false according to the geometry of Lobachevsky. There are many similar sophisms, both from ancient Greek and more recent ones. As for the second statement, it belongs neither to the humanities nor to philology, being very far from both fields of scholarship. So it is obvious that further discussion on the relationship of philology and mathematics needs to be continued on a firmer basis.

■□

I was not a student of Kolmogorov. As for my education, I am not a mathematician, I'm a philologist and a logician. But I so highly value his work in the field of analyzing poetic texts by means of mathematical statistics and probability theory, and he, being such an extraordinary and remarkable person, played such an important, even crucial

[1] In the translation of this article, the bibliography and titles of poems within the text are transliterated on the basis of the Library of Congress transliteration system without diacritics, while the transliteration system used for proper names in the text differs slightly from that in the bibliography. Within the text, proper names that normally end in –ii or –yi become simply –y (Tomashevskii becomes Tomashevsky). Quotations from poems are given in Cyrillic, as well as individual words in sections dealing with meter schemes and rhyme. This approach is used by Barry P. Scherr in [B. Scherr. Russian Poetry: Meter, Rhythm, and Rhyme. University of California Press, 1986]. *(Translator's note)*

role in my life, that I feel obliged to share everything I know about him, everything to which I was a witness. That is why, in awe and trepidation, I am proceeding to realize my plan: to impart my knowledge about Kolmogorov's studies in verse theory. This will be my humble contribution to kolmogorovedenie. And if this sinful servant of God hath written not, or hath confes'd too much, thou shalt not curse him in the presence of the Lord, but keep thy tongue from evil and correct him, such is the plea I address to my reader, following an Old Russian scribe[2].

■□

Verse theory demands the most profound study, for the 18th - 20th centuries marked the emergence and development of the great poetry by Pushkin, his predecessors, contemporaries and successors. Though this poetry was mainly based on the syllabo-tonic system, it also employed the tonic system, vers libre, the dol'nik, logaoedic verse and various imitations of folk verse. Alongside the Russian novel from the second half of the nineteenth century, this poetry became the most vivid expression of the spiritual quest of the Russian people and, I think, their greatest contribution to world culture.

In the middle and the second half of the twentieth century a few people created a mathematical theory of poetic speech. They included K. F. Taranovsky, academician A.N. Kolmogorov, A.V. Prokhorov, R.O. Jakobson, academician M.L. Gasparov, M.A. Krasnoperova, M.G. Tarlinskaya and the author of this essay. Our American colleague, Professor James Bailey, called this theory "the Russian method". Our work built upon the ideas of A. Bely, B. V. Tomashevsky, G.A. Shengeli, and M.P. Shtokmar. I do not claim that either part of this list is exhaustive but at the same time I see no reason to expand it. Our work was carried out at the time when nuclear physicists were creating more and more powerful hydrogen and super hydrogen bombs, while politicians were awkwardly juggling these bombs on the international political stage, like clowns juggling their balls in a circus ring.

Once, when I was sharing these ideas of mine with a well-educated engineer, who was something of a bigwig, he objected:

– How can you compare such things? What's the use of your mathematical theory of poetic speech?

– And what's the use of your hydrogen bombs? – I asked him.

My interlocutor, my former classmate, was struck dumb. He kept silent for some time, thinking it over. Then he said that he was not going to discuss my paradoxes and changed the subject to football. There had been a time when we played on the same football team.

■□

I distinguish three main aspects of Kolmogorov's activities in the field of verse study.

1. The problem of the axiomatic construction of a theory for Russian classical metrics.

2. The study of Russian verse meters outside classical metrics.

3. Supporting the use of exact methods in literary study.

[2] In the original Russian text, the author of the article, V.S. Baevskii, uses the Old Church Slavonic variant of the language to create a solemn and sacred aura: «Еже азъ многогрешный недописахъ или переписахъ, не кляни Бога для, а возьми да поправь, говорю я читателю вслед за древнерусским книжником». *(Translator's note)*

1. The problem of the axiomatic construction of a theory for Russian classical metrics

Probability theory emerged in the seventeenth century. Pascal and Fermat established probability theory, or the so-called mathematics of random phenomena, primarily by watching people playing cards and dice. Most people, if they do not have complete knowledge about a subject, tend to think that they know nothing about it. But it turned out that even incomplete knowledge allows us to measure the degree, or the quantity, of knowledge. Total lack of knowledge is conventionally assigned a 0, while complete knowledge is 1. The probability that any random event will occur lies between 0 and 1. Probability theory cannot be applied to nonrecurrent random events. Sometimes even mathematicians do not take this into consideration. E.S. Ventzel, an outstanding expert on probability theory, once related how she had been amazed when she dropped in at a session being held at a mathematics institute. The author of the report was dwelling on the probability that the Baikal-Amur Railway would be constructed. Nobody even reminded him that this theory can be applied only to the study of repeated homogeneous phenomena.

Originally, the probability of any random event occurring was predicted on the basis of empirical observations. Mathematicians used to say: "The probability of any random event occurring under certain conditions is what it generally was in the past under similar conditions."

Kolmogorov's objective was to construct probability theory as other branches of mathematics are constructed – axiomatically, deriving its primary propositions from a few assertions that rest upon the fundamentals of mathematics and logic. At the end of the 1920s and the beginning of the 1930s, when Kolmogorov was about thirty years old, he published some essential works on this topic in German and French. A few years later these works were published in Russian. Some people feel that Kolmogorov had fully attained his goal, others hold the opinion that he did not succeed in completing this theory. I have no right to judge. However, everyone agrees that he went far along his projected path and laid the groundwork for the axioms of probability theory. Nowadays any specialist in probability theory can say: "I am able to predict the probability of the emergence of any random event within fixed limits because my prediction is based on the foundations of mathematics and logic."

The decades of Stalinism were marked by the suppression of all scientific disciplines that used exact methods to explore the spiritual world of a human being. Sometimes such suppression had a disastrous impact on the scientists' lives.

Until the middle of the 1950s, i.e. until Stalin's death, any extensive research on poetic speech with the help of exact methods was impossible, since it was considered to be formalism. In its turn, formalism was virtually equaled to anti-Soviet activity. Thus, a scholar was sentenced to death because he had belonged to the formalist school in his youth. Another had to renounce the formal method publicly in order to save his life, as Galileo did long ago.

From the middle of the 1950s to the beginning of the 1980s, Kolmogorov and his students were trying to construct an axiomatic theory of poetic speech – these were practically thirty years of more or less systematic work on the application of probability theory and mathematical statistics to the study of poetic speech. Illness and, finally, death put an end to this work.

Almost ten years after Stalin's death, papers by Kolmogorov and his disciples on verse theory began to appear. Eight works out of eleven were published during the relatively liberal 1960s — to be more exact, in 1962–1968.

I know the following articles by Kolmogorov dealing with the problems of verse theory: (Kolmogorov 1968: 5–14). Each paper contains the analysis of one or two verse forms, and three of them [Kolmogorov, Rychkova 1999; Kolmogorov, Prokhorov 1968, 1985) clearly reflect the author's aspiration (which unites all his studies) to work toward a general outline of the axiomatic theory of poetic speech.

The works of Kolmogorov and his students show that he saw no point in getting into a detailed analysis of items without working out a general approach to the problem. At the same time he saw no point in any general discussions which are not based on the analyses of poetic texts at the most serious level available hic et nunc.

In accordance with the tradition generally accepted in verse study, Kolmogorov used the term 'meter' in two different meanings: 1) "meters of Russian syllabo-tonic versification – trochee, iamb, dactyl, amphibrach, anapest" and 2) "the meter of a given text – the whole set of restrictions imposed on the poetic speech of a given text." However, sometimes there may be no clear-cut boundaries between these two definitions, because the meter of a given text is primarily dependent on the rules of the corresponding meter for all Russian poetry, and only after that does it impose special restrictions. The aggregate of the most significant restrictions is commonly referred to as a measure: trochaic trimeter, iambic pentameter with a caesura after the second foot, etc.

Kolmogorov was trying to develop an axiomatic theory of poetic speech in order to reveal and describe the fundamental propositions on which verse theory is based. In his published works he never used this very term, "axiomatic theory," speaking only about methods of research, its rules and laws. Nevertheless, close study of his papers on probability theory, logic and verse theory has made me think that in all these cases he was expressing this same idea of trying to establish an axiomatic theory.

Kolmogorov studied various versification systems but his attention was mostly directed to the syllabo-tonic system – the most wide-spread and predominant in Russian versification.

First of all, he considered it necessary to elucidate all the restrictions that the most widely used syllabo-tonic meters impose on contemporary prose speech. After that he intended to generalize these results— i.e. to single out those restrictions imposed on prose speech that are characteristic of the entire syllabo-tonic system.

A detailed presentation of such a generalization first appeared in an article written together with Prokhorov (Kolmogorov, Prokhorov 1968). The final formulation of this concept—a generalization containing four fundamental propositions—was set forth in a subsequent article by Kolmogorov and Prokhorov (Kolmogorov, Prokhorov 1985: 114). The authors refer to these propositions as laws, and reduce them to the following:

1. Each verse (line) contains a certain number of syllables which is specific to a given metrical scheme.

2. A metrically strong syllable (ictus) is likely to be stressed (though the stress may be omitted).

3. A word-boundary obligatory for the metrical scheme (for instance, a caesura after the second foot in iambic pentameter) can be found only between two different rhythmic words. This may seem to be a tautology but in fact it is not. From the point of

view of versification, a rhythmic word includes not only the word itself but any proclitics and enclitics as well. Thus, the verse *// Пусть светит // месяц // – ночь // темна* contains four rhythmical words: *Пусть светит; месяц; ночь; темна.* The interjection *пусть* cannot be considered as an independent rhythmic word; it is adjoined to the notional part of speech – the verb *светит.* There is no word-boundary between these two words. The verse *Ты проснешься ль, // исполненный // сил* contains three rhythmic words: *ты проснешься ль* (a notional part of speech – the verb *проснешься*, a proclitic – the pronoun *ты*, an enclitic – the particle *ль*); *исполненный* and *сил*.

4. A weak syllable may have a stress only when none of the word's other syllables fall on the strong position (the ictus). For instance, in the verse *Швед, русский колет, рубит, режет* the stressed word *швед* may be in a weak position (an odd iambic syllable) because it is a one-syllable word, with no syllables on an ictus. In "The Lay of Opanas" ("Duma pro Opanasa") this fourth rule is not observed. Imitating Ukrainian folk verse, an example of which is given by the author in the form of an epigraph, Bagritsky writes:

> По откосам виноградник
> Хлопочет листвою <…> (Bagritskii 1934: 73)

The second line does not adhere to the trochaic meter: the stress of the first word is not on an odd but on the second, even, syllable. Moreover, the first, unstressed syllable of this word is on an ictus. In this way Bagritsky achieves a specific semantic and artistic effect: from the very beginning the author shows that the poem is about the events in the Ukraine, about the destiny of the Ukrainian people.

■□

In order to formulate these four laws Kolmogorov had to carry out a series of investigations devoted to Russian verse meters, which, independent of their importance for devising these laws, play a significant role in understanding both Russian poetic speech and Russian poetry in general. Let us now discuss some results of these investigations.

Kolmogorov deepened the understanding of verse meter. He advanced the opinion that "in the living perception of a poet and a listener meter does not exist as a kind of mere rule <…> but as a specific artistic image" (Kolmogorov, Prokhorov 1963: 84). He was sure that a poet's work leads to the "creation of a rhythmic image that is cast into the crystalline form of meter" (Kolmogorov 1968: 147). Poets themselves talk of this from time to time. Let us recall one such episode.

Shortly before his death, Blok was going to take part in a ceremonial meeting dedicated to the 84th anniversary of Pushkin's death. He wrote a wise and harmonious prose poem in his diary: "What is a poet? – A person who writes verse? Of course not. The poet – the poet is the bearer of rhythm. In the endless depths of the human spirit, in depths that are beyond the reach of anything earthborn, disturbed neither by morality, nor by law, society, or government – sound waves, related to the waves embracing the universe, are vibrating; rhythmic oscillation takes place there, similar to the oscillation of heavenly bodies, glaciers, seas, and volcanoes" (Blok 1963: 404–405). After modifying it slightly, Blok included this text in his speech delivered at that meeting. A study devoted to this prose poem would show that its main idea was inspired by academic literary scholarhsip from Buslaev to Veselovsky. Here I cannot help but

mention that Blok's handful of philological works demonstrate that the brilliant poet was also an outstanding scholar and literary researcher. One of the greatest psychiatrists of the twentieth century wrote: "The most primitive rhythmic tendencies, though crowded out by purposeful activities and hidden behind more complicated sensations, are still living in the lowest layer of our psychomotor apparatus <…> A simple rhythm, as well as optical symmetry, still awakens in us a familiar sense of pleasure, which has its deepest roots in the very depths of phylogenetic development and cannot be reduced further" (Krechmer 1927: 100).

Blok anticipated the important ideas and rhythms of his prose poem in the preface to his unfinished poem "Retribution" ("Vozmezdie"). The preface was written two years earlier and contained "the story of how the poem was created, of the causes of its emergence, and the origin of its rhythms". In the preface he listed the most varied facts illustrating Russia's life in 1910 and the adjacent years – the deaths of Kommissarzhevskaia, Leo Tolstoy and Vrubel', the crisis of symbolism, revolutionary currents, a premonition of the World War, the Beilis case, the rise in popularity of free-style wrestling, the development of aviation, the murder of Stolypin. He drew a general conclusion: "Though all these facts may seem to be so different from each other, for me (the poet! – *V. B.*[3]) they are united by one and the same musical meaning <…> I think that the simplest expression of the rhythm characterizing that time, – a time when the world, preparing itself for incredible events, was earnestly and systematically developing its physical, political and military muscles, – was the iamb. Probably I was no exception, driven over the world and being whipped by this iamb, I finally had to surrender and give in to its springy wave for a prolonged time." (Blok 1960: 295–297)

Kolmogorov distinguishes two aspects of verse meter: 1) the sound image; 2) its semantic interpretation. (From the point of view of semiotics, the sound image is a significant, the semantic interpretation is a designatum.) The sound image is made up of two kinds of phenomena. The nature of the first kind is determined, and its manifestation is inevitable. For instance, in Lermontov's ballad "Tamara" the second, the fifth and the eighth syllabic positions of each verse have obligatorily stressed syllables. The second kind is probabilistic in nature and is manifested through the prevalence or weakening of certain tendencies. One may frequently observe a great stability of average values; for example, the percentage of stressed ictuses in trochaic and iambic tri-, tetra- and pentameter in the poetry belonging to one or a number of poets throughout two or three decades of their work. "An adequate way of describing the sound image for a meter is the statistical data showing the use of rhythmic types in individual lines and their combination in successive lines. The research carried out by Bely, Tomashevsky, Shengeli, and Taranovsky," Kolmogorov writes, "has shown that the sound image of the meter in individual large works (see the classical studies of *Eugene Onegin* by Tomashevsky), among individual poets, or among poets of the same literary movement frequently display high degree of consistency." (Kolmogorov, Prokhorov 1963: 85).

Like Tomashevsky, Kolmogorov attributed the highest aesthetic significance to those forms and rhythmic variants that more or less significantly deviate in their frequency from the average speech values. That is why, when analyzing meter, great importance is given to juxtaposing the results that were obtained through comparing

[3] Initials of the author, V. Baevskii *(Translator's note)*

the data for a given text with the theoretical probability model for that phenomenon. Let us take as an example the analysis of Pushkin's "Arion" made by Kolmogorov. Since the text of the poem is easy to find, we will not quote it here.

This is a rather complicated study, which contains a thorough examination of Pushkin's work at various levels and in different aspects. We shall treat here only Kolmogorov's main idea. He pays attention to the fact that the poem may be divided into three parts according to its meaning, its syntactic structure, its rhyme system and its rhythm (the poem is composed in iambic tetrameter). The first part contains four lines, the second – five, and the third – six lines. All four lines of the first part, which depicts a peaceful voyage, belong to the most commonly used rhythmic form in Russian poetry in general and in Pushkin's poems in particular: the form of iambic tetrameter with stresses on the first, the second, and fourth ictuses, and with an unstressed third ictus. The second part tells the reader how this peaceful voyage was interrupted by a dangerous storm, and here the iambic rhythm changes accordingly. All five lines of this part are fully stressed, and all the ictuses are occupied by stressed syllables. In addition, there is one more stressed syllable, which falls on a non-ictic position. This is the adverb 'вдруг' marking the peripeteia – a sudden turn from a peaceful voyage to a storm. The rhythm of the third part of the poem becomes the same as it was in the first part, which exactly corresponds to the main message of the entire text: the peaceful voyage was interrupted by a storm, the sailors perished and only the minstrel has survived and keeps singing the former songs. And the symbol of these former songs is the former rhythm. If we concur that the poem, written soon after the Decembrist revolt and Pushkin's release from exile, reflects his thoughts about his own destiny as set against the fates of the Decembrists, among whom were his close friends, and of Russia as a whole, then the metrical image created by the poet becomes especially expressive and meaningful.

One should not think that the difference in the rhythm between the middle, second part and that of the first and third parts occurs just by chance. Kolmogorov provides calculations showing that the chances of having a quatrain and a sextain with stresses on the first, second and fourth ictuses, unless the author should do this consciously, are close to 0.5%. As for the probability of a quintain with all the ictuses stressed (again without the author doing so on purpose), it is close to 0.001—i.e. extremely low.

In Kolmogorov's terms, one would say that the essence of the metrical sound image in "Arion" is a three-part division of the text, with the first and the third parts contrasted to the second. The essence of the semantic interpretation of this metrical sound image is the notion of the author's adherence to the ideas of his Decembrist friends after the suppression of their revolt.

This rhythm that Kolmogorov discovered for the three parts of the poem only approximately corresponds to the above-mentioned development of the ideas. The peaceful voyage is described in the first seven lines, not four; the storm is mentioned in three lines; the destiny of the minstrel who survived and is singing old hymns is told in the final five lines. Some discrepancy between the boundaries of these three semantically and rhythmically determined parts of the poem is natural: the poet maps out the general pattern of the rhythmic development, but does not force a one-to-one correspondence between the rhythm and the meaning. He is an artist, not an engineer.

I do not want the reader to think that the rhythmic expressiveness of "Arion" is exceptional. Therefore I'll provide one more example from the works of another poet,

who is quite different from Pushkin – “Let me bring forth a poem in suffering!..” (“Dai vystradat’ stikhotvoren’ie!..”) by David Samoilov.

Дай выстрадать стихотворенье!
Дай вышагать его! Потом,
Как потрясенное растенье,
Я буду шелестеть листом.

Я только завтра буду мастер
И только завтра я пойму,
Какое привалило счастье
Глупцу, шуту, Бог весть кому, –

Большую повесть поколенья
Шептать, нащупывая звук,
Шептать, дрожа от изумленья
И слезы слизывая с губ.

This poem was written 22 July 1967, published in the magazine *Novyi mir* (no. 2, 1969), reprinted in Samoilov’s books *Days* (*Dni*) (Moscow, 1970) and *Equinox* (*Ravnodenstvie*) (Moscow, 1972) and in collections of his verse that appeared both during his lifetime and posthumously. We are again dealing with iambic tetrameter, but the rhythm of the first quatrain is entirely different from that in “Arion”.

Below we shall be making use of the data in Table III, which is an appendix to Kolmogorov’s classical work (Taranovskii 1953). Using Tynyanov’s term that was created to designate a group of original poets from the first half of the nineteenth century, we shall refer to Samoilov as an “archaizing innovator” from the second half of the twentieth century. Therefore, it makes the most sense to compare the rhythm of his iambic tetrameter with the data for the iambic tetrameter of Pushkin and his contemporaries.

From the point of view of rhythm, the first line is a rarity. An iambic tetrameter line with stresses on the first and fourth ictuses occurs with an average frequency of 0.004. And a line in this rhythmic form that has a hypermetrical stress on the first, non-ictic syllable is doubly rare. The second line, with stresses on the first, third and fourth ictuses, cannot be called rare though it also is not common. Its frequency is 0.07. As in the first line, Samoilov introduces a hypermetrical stress on the first, non-ictic syllable in order to make it sound exceptional. Such lines, with stress on the first, third and the fourth ictuses together with hypermetrical stressing on the first non-ictus, are extremely rare. But the poet does not stop his experiments here. He creates a syntactic pause in this line; the combination of all these devices makes the second line unique. The rhythmic form of the third line, with stress on the second and the fourth ictuses, is not among the more frequent (its frequency fluctuates around 0.08) but it is far from being as exceptional as the rhythmic forms of the two first lines. And finally, the fourth line, which concludes the first quatrain, like the second line, has the most common of the rare rhythmic forms (0.07).

The rhythm in the second quatrain changes beyond recognition. The first, second and fourth lines belong to the fully-stressed rhythmic form that is already familiar to us from the second part of “Arion.” The frequency of this form in the texts used for

comparison is 0.31. The second ictus of the third line lacks stress. Here we have a so-called "change in the third quarter", an aesthetic device that is rather well known to musicians, both composers and performers. In general, the rhythm of the second quatrain is much simpler than that of the first quatrain.

The rhythm of the third quatrain displays the most frequent rhythmic form: all the lines have stress on the first, second and fourth ictuses (as in the first and the third parts of "Arion"). Their frequency in our comparative material is 0.46.

So, over the course of the poem the rhythm of the lines gradually changes, going from the most complicated form to the simplest.

Originally there was one more quatrain between the first and the second stanzas:

И буду шевелить губами
Я, изумленный Нострадам,
Все, что начертано судьбами,
Прочитывая по складам. (Samoilov 2006: 595)

But then Samoilov crossed it out. While it would be difficult to claim that we know the exact reason for any change a poet makes to his text, in this case I have some thoughts I would like to share.

"It's the composition that really matters," Samoilov once told me. "It all lies in the composition. Ninety per cent of the poems in [the anthology] *Poetry Day* (*Den' poezii*) lack composition. You could shorten them or lengthen them. Out of twenty lines you could leave only eight. As for me: when the poem was completed I would cut out a good half of the stanzas. What's the use in telling absolutely everything? It's much more enjoyable for the reader to interpret it by himself. Only in that case will he perceive this poem as part of his own life experience."

The rhythm of the stanza cut out by Samoilov, is almost as complex as that of the first quatrain. Two lines are characterized by hypermetrical stress on the first non-ictus; enjambment between the first and the second lines complicates the rhythm; none of the lines represents any of the more common rhythmmic forms. In regard to the rhythm going from complicated to simple, the "eliminated" quatrain seems to be superfluous: it delays this movement.

This change of rhythm from complicated to simple supports the semantics of other levels and aspects of the poem (see Baevskii 1976, 1982). The meaning of the poem may be briefly conveyed as follows. The poet feels that he is chosen to say the most important things but is unable to do so. A huge creative force has been accumulated in the depth of his being and now is demanding to come out. He prays to some mysterious highest authority whose name he most likely does not even know – God? Apollo? Muse? – to allow him to release that creative force inside him (the first quatrain). His previous experience tells him that only after this discharge has occurred will he come to realize the deepest meaning of what this higher power has created through him (the second quatrain). Happy and exhausted, he will understand that he has created a great story of his generation (the third quatrain). The verse rhythm, which starts with maximum tension and goes to maximum slackness, accompanies the poet's spiritual path and in combination with other devices creates the great story of the generation which is the subject of the poem.

■□

2. Studies of Russian verse outside the syllabo-tonic system

Apart from the iambic tetrameter, Kolmogorov, both independently and together with Prokhorov and Kondratov, published studies on the versification of Pushkin's "Songs of the Western Slavs" as well as on the verse of Akhmatova, Tsvetaeva, Maiakovsky and Bagritsky (Kolmogorov 1968: 6–11). This article, according to Kolmogorov, is an introductory chapter from his unfinished textbook on verse study for beginners. As a first exercise the author suggested analyzing a simple verse meter from Tsvetaeva's cycle "Desk" ("Stol") (poems 1 to 5) and her "Poem of the End" ("Poema Kontsa") (the fifth, sixth and ninth parts). Those around Kolmogorov often wrote and said that he was extremely difficult to listen to. He seemed to think that the world was populated by Kolmogorovs and not by ordinary people. That is why it was very difficult for him to make his narration accessible to others, to adjust himself to his listeners' level of understanding. Occasionally, one would come across such paradoxes as the following: "The lecture of Kolmogorov was incomprehensible but interesting". Personally I remember one lecture on the basics of Russian classical metrics, delivered by Kolmogorov in February 1970 to L.I. Timofeev's symposium on verse study at the Institute of World Literature of the Russian Academy of Science. I knew all his publications on this subject and several papers on mathematical logic, so I understood the lecture without, I believe, any lacunae. (There is one other lecture by Kolmogorov that is also quite vivid in my mind. I understood all of it without difficulty and, I think, for the same reason: I had studied the works on information theory written by Kolmogorov, Ashby and some others that were relevant for the matters being discussed.) But I am not going to recount Kolmogorov's article about his analysis of Tsvetaeva's verse. Those who are interested may read the original. My task is to describe the main line of Kolmogorov's thought in carrying out this analysis.

When Kolmogorov published his article, Tsvetaeva's name could be mentioned in public but it was strictly forbidden to praise her. Brodsky's name was under a ban. Meanwhile, people who belonged to literary circles were already aware that it was Tsvetaeva who, in Brodsky's opinion, was the best Russian poet of the twentieth century. Kolmogorov could not have failed to know this.

The meter of "Desk"[4] and "Poem of the End" has no generally accepted name in the scholarship on Russian verse. It is a transitional case between the dol'nik and logaoedic verse.

The *dol'nik* is verse in which either one or two weak syllabic positions may appear between two strong syllabic positions. And that's the way Tsvetaeva wrote these texts.

The strong syllabic position, or ictus (Latin: 'ictus' – 'a beat, a pulse, a word stress'), may be filled by either a stressed or an unstressed syllable.

The weak syllabic position, or non-ictus, may be filled only by an unstressed syllable. This rule has one exception: in a weak syllabic position (non-ictus) there may be a monosyllabic notional or semi-notional word that bears stress.

[4] "Desk" is one of the best-known poems by M. Tsvetaeva. Here we give an English translation of the most frequently quoted quatrain: *My desk, most loyal friend // thank you. You've been with me // on every road I've taken. // My scar and my protection*. Trans. by Elaine Fernstein. *(Translator's note)*

The question of semi-notional words arose in Russian verse theory about a century ago; it was Andrei Bely who first touched upon this problem. Briefly, the issue is that such semi-notional words as pronouns, prepositions, conjunctions, interjections and particles may be perceived as stressed, or as having stressed syllables if they are bi-syllabic, but they may also be perceived as unstressed.

In logaoedic verse the number of non-ictuses between the strong syllabic positions in the line varies, but the ictuses and non-ictuses are located in the same positions within every line of such a poem. And that's the way Tsvetaeva wrote these texts.

We may say that Tsvetaeva's verse in these works consists of a "congealed" dol'nik — or logaoedics in the form of a three-ictus dol'nik. The first syllabic position is a non-ictus, the second is an ictus, the third and fourth are non-ictuses, the fifth an ictus, the sixth a non-ictus, and the seventh an ictus. There may or may not also be an eighth position; its designation in a metrical scheme is conventionally given in parentheses. It is always a non-ictus.

This is the distribution scheme of ictuses and non-ictuses in the meter by Tsvetaeva that Kolmogorov examined:

n i n n i n i (n) Мой письменный верный стол!
Я знаю твои морщины

This is the *first* level of analysis: Kolmogorov establishes the distribution of ictuses and non-ictuses in all the lines of the text that is being analyzed. He explains that he has chosen this text because it is rather simple to study.

The *second* level of analysis reveals how the ictuses are stressed. In Russian versification the final ictus of a line carries an obligatory stress. When that syllable is the last in the line, the clasusula is masculine. The final ictus may be followed by one non-ictus (feminine clausula), two non-ictuses (dactylic calausula) or more than two non-ictuses (hyperdactylic clausula). These texts by Tsvetaeva contain only masculine and feminine clausulae.

It is easy to see that there are 4 ways of stressing the ictuses (4 rhythmic forms) in the text:

I	n iS n n iS n iS (n)	Со мною по всем путям
II	n i n n iS n iS (n)	Не похоронив – смеяться!
III	n iS n n i n iS (n)	Все заповеди Синая
IV	n i n n i n iS (n)	<А из-за нерасторопной>

Tsvetaeva's text does not containthe fourth rhythmic form. Here we have provided an artificially composed example. Kolmogorov explains that such long words are very uncommon in the Russian language, and Tsvetaeva avoids using them in her poems, whatever the meter (Kolmogorov 1968: 148).

The *third* level of analysis reveals the stresses on non-ictuses. However, as mentioned above, the word must be monosyllabic and notional or semi-notional. Kolmogorov distinguishes only the three forms with stress on the first non-ictus, the anacrusis; these are variants of the first three rhythmic forms, respectively: V stands for the first rhythmic form with hypermetrical stress on the first non-ictus, VI indicates the second form with hypermetrical stress on the first non-ictus, and VII represents the third

form with hypermetrical stress on the first non-ictus. Kolmogorov called these variations of the first three rhythmic forms.

V nS iS n n iS n iS (n) Стол – вечный – на весь мой век!
VI nS i n n iS n iS (n) Смерть – и никаких устройств!
VII nS iS n n i n iS (n) Стол – сбрасывавший – в поток!

Kolmogorov studies only metrical patterns with a stress on the first non-ictus, the anacrusis, whereas Tsvetaeva's rhythm is much more complicated. For example, she has the following line:

Бог! *Есть* Бог! Поэт – устройчив (Tsvetaeva 1965: 301)

All the rules typical for Tsvetaeva are observed in this line: there are 7 syllables from the first non-ictus to the final ictus; all the ictuses are stressed; the verb *'есть'* in the second syllable (the first ictus) was made prominent by Tsvetaeva herself, in order to show the verse rhythm more clearly:

бог ЕСТЬ бог поЭТ уСТРОЙчив

One can clearly see the three stressed ictuses as well as the non-ictuses in the first, third, fourth and sixth positions; this scheme is also observed in all the other verse lines.

However, this line is characterized not only by stress on the first non-ictus (the first syllable) but also by stress on the second non-ictus (the third syllable). So it is highly probable that Kolmogorov, in order to make the analysis easier, examined only the variants with stress on the first non-ictus, the anacrusis. Otherwise, his analysis would have become much more cumbersome.

This three-level study of her verse enabled Kolmogorov to reveal the law which had caused Tsvetaeva to select, out of the unlimited amount of expressions that are possible according to the rules of Russian grammar and within the Russian lexicon, those that comprised five poems in her "Desk" cycle and three parts of "Poem of the End".

But meter arises and exists in the poet's consciousness and Subconsciousness not only in the form of a law. At this point the *fourth* level of metrical analysis begins. Kolmogorov has established that meter exists in the poet's consciousness and Subconsciousness as a pulsatory rhythm of varying repetitions of the different rhythmic patterns. To begin with, I should say that frequently used patterns serve to create an emotional background, a relatively calm narration, whereas the rare rhythmic forms help emphasize certain lines and line groups in peripeteias, different types of crises, catastrophes, and closure. Kolmogorov draws the reader's attention to the fact that "in the first poem of the 'Desk' cycle Tsvetaeva, like a careful nanny, helps the reader grasp her meter's possibilities. The first two stanzas show the meter in its pure form" (Kolmogorov 1968: 160). All eight lines of these two quatrains belong to rhythmic form I: each ictus is stressed, and all the non-ictuses are unstressed. If we examine the closing quatrain of the first poem in the "Desk" cycle, we shall fail to find any line using rhythmic form I. In this quatrain forms III and V alternate.

■□

Kolmogorov lists the results of his counts of the various rhythmic forms and variations in the verse of "Desk" and "Poem of the End". Here we shall consider the data on the "Desk" cycle. Rhythmic form I is found in 98 lines, there are no examples of form II, form III occurs in 14 lines, V in 11 lines, VI in 1 line, and VII in 4 lines. Altogether 128 lines were examined. Kolmogorov expresses the ratio of these forms as a percentage: I – 76.6; II – 0.0; III – 10.9; V – 8.6; VI – 0.8; VII – 3.1.

As stated above, the rare rhythmic forms often mark peripeteias, various kinds of crises, catastrophes, and closure, usually concentrating in themselves all the creative energy spent on the entire poem. Let us see which lines these rhythmic italics single out in "Desk." The first part contains a quatrain with two lines that belong to rhythmic form VII:

Всем низостям – наотрез!	nS iS n n i n iS	rhythmic form VII
Дубовый противовес	n iS n n i n iS	rhythmic form III
Льву ненависти, слону	nS iS n n i n iS	rhythmic form VII
Обиды – всему, всему.	n iS n n iS n iS	rhythmic form I

In the "Desk" cycle Tsvetaeva opposes herself, a poet in exile living for spiritual values and creative work, to commonplace people who hated her and knew nothing but grubbing – in this respect, the final poem of the cycle is particularly important, though Kolmogorov did not include it in his analysis because it was composed in a different meter. This opposition is clearly expressed in the above quatrain, where the rhythmic italics reveal her opposition to meanness and to the "lion of hatred."

Four lines bearing rhythmic italics appear at the end of this first poem in the cycle. Different hymns to her desk are sung in each line: *стол, выстроивший в столбцы <…>*; *столп столпника, уст затвор –; лбом, локтем, узлом колен* – and finally comes the closing line of this poem – *В грудь въевшийся – край стола!*

The reader, if he wishes, may examine the role of rhythmic italics further: *Стол – сбрасывающий – в поток!* (the second poem); *Да, был человек возлюблен! Ты – стоя, в упор, я – спину <…>, Бог! Есть Бог! Поэт – устройчив: Все – стол ему, все – престол! Ты – мой наколенный стол!* (third poem of the cycle; it too concludes with an italic line); *Дал, стройный, врагам на страх – Стол – на четырех ногах <,,,> Стол – вечный – на весь мой век!* (fourth poem). There are no rhythmic italics in the cycle's fifth poem. Both quatrains comprising it are written in rhythmic form I, as was the case in the cycle's two initial quatrains. The result is a compositional rhythmic circle.

■□

To grasp the message of Tsvetaeva's rhythmic italics we have to touch upon the semantic aspect of the "Desk" cycle. As we once observed many years ago, the lyrics of Tsvetaeva's mature period are to a great extent based on the device of semantic variation. An idea arises in Tsvetaeva's mind, usually in a metaphorical and rhythmic form, in the form of an aphoristic formula. This embryo of a future poem becomes its invariant. In a way it resembles the work's pathos, in Aristotle's and Hegel's meaning of the word. When realized in a text, the invariant undergoes multiple changes, and these variations make up the basic essence of the text. One poem may have several invariants, interwoven throughout the text and making its message far more complicated (Baevskii 2003: 172–173).

The first line of the cycle's initial poem presents one of the invariants, from the point of view of both its meaning and rhythm. It is a direct address: *Мой письменный верный стол! (My desk, most loyal friend…).* It will be repeated exactly at the beginning of the final, fifth poem, and in variations throughout the cycle over and over again:

Мой письменный вьючный мул!
Строжайшее из зерцал!
Мой заживо смертный тес!
Деяний моих столбец!
Столп столпника, уст затвор!
В грудь въевшийся – край стола!

Но лучше всего, всех стойче –
Ты, – мой наколенный стол!

Стол – вечный – на весь мой век!

After addressing her desk, Tsvetaeva expresses her gratitude to it in multiple and diverse variations:

Спасибо за то, что шел
Со мною по всем путям.
Меня охранял, как шрам.

Спасибо, что ног не гнул
Под ношей, поклажу грез –
Спасибо – что нес и нес.

Спасибо за то, что стал
(Соблазнам мирским порог),
Всем радостям поперек,

Всем низостям – наотрез!

The last line belongs to the next quatrain.

We won't be quoting further variants of the gratitude Tsvetaeva expresses to her desk, and will show variants of other invariants; we shall give a few examples. Right before our eyes the invariant of gratitude has developed into a description of the temptations against which the desk guards its master. Then this invariant is developed further, nearly throughout the entire cycle:

Спасибо, что рос и рос
Со мною, по мере дел
Настольных – большал, ширел <…>

And so forth, constantly growing. In the same way she develops the invariant of the thirtieth anniversary of the alliance between the poet and her writing desk — i.e., the

thirtieth anniversary of her creative work: *Я знаю твои морщины (I know your wrinkles)*. This grows into a hymn to the desk, essentially a unique hymn glorifying the tragic fate and creative work of the poet. After the middle of the cycle one more invariant is introduced: "*труд поэта угоден Богу*" *("poet's work is pleasing to God").*

Other, less clearly delineated invariants also appear now and then. Here we can pause and see to what extent each of the above-mentioned basic invariants, essential for the poem's texture, is marked with rhythmic italics – forms VI and VII.

The address to the desk is not marked by rhythmic italics.

Gratitude is marked once with rhythmic form VI and twice with form VII:

Стол – на четырех ногах	form VI
Льву ненависти, слону	form VII
Стол, выстроивший в столбцы	form VII

The invariant "*Труд поэта угоден Богу*" is marked the most strongly: by a verse line which has three stresses on the ictuses and two stresses on non-ictuses and therefore does not fit within Kolmogorov's nomenclature of possible variations.

Бог! *Есть* Бог! Поэт – устройчив nS iS nS n iS n i S n

Besides the rhythmic italics, this line is also notable for some other devices: enjambment which links this verse with the preceding quatrain; the italics, which emphasize the verb *есть*; two exclamation marks; three pauses (after the two exclamation marks and the dash); and the nonce word *устройчив*. This line is the most notable among all 128 lines in the first to the fifth poems of the cycle. So, there are two semantic invariants and their variations which are rhythmically marked: the "*труд поэта угоден Богу*" type and the gratitude to the writing desk for the gift of being a poet. Gratitude to the desk is gratitude to God. "*Спасибо Тебе, Столяр*" – Tsvetaeva writes at the end of the fourth poem. Thus, the sound image of the meter, through its semantic interpretation, helps the poet express the main pathos of the lyric cycle.

■□

The *fifth* level of Kolmogorov's research touches upon the problem of the relationship between poetic speech and natural language. Tomashevsky was the first to bring up this question for discussion (Tomashevskii 1971; 1929. 104–105), and Kolmogorov elaborated on some of his points. If I am not mistaken, he never provided a detailed description for this part of his method in his published work; it is known only from his students' papers. He mentions it briefly in his paper, which we have just discussed, on Tsvetaeva's verse (Kolmogorov 1968: 152). The same collection of works in which Kolmogorov's article was published contains a monographic paper by M.L. Gasparov. The latter provides a thorough description of the methodology for using a probability-based theoretical model of a verse meter. After it was published, this paper received some criticism; therefore it must be treated somewhat carefully. For instance, the model that Gasparov regards as reflecting Tsvetaeva's individual style (p.102), does not correspond to the basic metrical pattern of the "Desk" cycle. Nevertheless, his paper gives a good general idea of the issue. Here we shall explain how Kolmogorov's probability-based theoretical rhythmic pattern can be applied in his research on Tsvetaeva's verse.

Speaking somewhat theoretically, a poet, twice faces restrictions in the choice of words while composing a poem: first, when selecting the appropriate word from the language's lexicon and, second, when placing it into verse.

Language possesses words of quite varied rhythmic types—i.e. words having a different number of syllables and with different placement of stress. The study of frequency dictionaries, both of the language itself and of particular authors, shows that words of different rhythmic types are characterized by a different frequency of occurrence in Russian speech. For instance, there is a tendency to use short words more often than long ones, and the stress in Russian words is not confined to a specific position but tends to be drawn to the middle of a word. Tsvetaeva called her cycle "Desk". It is not clear whether she might have called it "Chiffonier" if the piece of furniture about which she had sung, had the three-syllable name "chiffonier." On the other hand, in the everyday life of a poet a desk is used much more often than a chiffonier. Consequently, it is more likely to be frequently mentioned. That is why the cycle is named with a short native Russian word and not a longer French borrowing.

So, a word with the needed meaning and of an appropriate length has been selected. Now it must fit within the line. The verse line has to have a certain number of syllables, which are to be stressed or unstressed according to the requirements of the given metrical pattern. If Tsvetaeva in her "Desk" cycle wants to make a line prominent with the help of rhythmic italics, she should stress the first non-ictus and select for that purpose a notional or semi-notional monosyllabic word. And if she has to show by means of rhythm which meter she has chosen, then she must choose three words that contain 7 syllables in total, from the first non-ictus to the third ictus. Moreover, the stresses must fall on the second, fifth and seventh syllables. Such are the restraints imposed upon a poet! In this case the following variations of rhythmic form I are possible, depending on the rhythmic types of words.

n iS ‖ n n iS ‖ n iS
n iS n ‖ n iS ‖ n iS
n iS n n ‖ iS ‖ n iS

n iS ‖ n n iS n ‖ iS
n iS n ‖ n iS n ‖ iS
n iS n n ‖ iS n ‖ iS

The reader may easily find in Tsvetaeva's lyrics examples of all these rhythmic words, from monosyllabic to four-syllable ones, and will experience great aesthetic pleasure in doing so. There were few poets who had as wonderful a command of Russian language and verse as did Tsvetaeva. Lest the reader think that she selected words and arranged them into rhythmic forms without thinking, "purely intuitively," note what she writes about the word "расстаемся" ("we are parting"), which is one of the key words in her poem dedicated to parting with her beloved:

Просто слово в четыре слога,
За которыми пустота (Tsvetaeva 1965: 467). (This is just a word of four syllables, beyond which there's only emptiness).

So, it turns out that Tsvetaeva counts syllables. Not without reason had she called her poetic activity work.

Nevertheless, in her "Desk" cycle there is no rhythmic form II without a stressed non-ictus and there is only one line with a stress on the first non-ictus in this form. This is how that form must look like without a stress on that first syllable:

n i n n iS ‖ n iS
n i n n iS n ‖ iS

Rhythmic form II without stress on the first non-ictus does not allow any other variants either in the placement of word boundaries or in the rhythmic types of words. Tsvetaeva would have had to find a five-syllable word with stress on the final syllable or a six-syllable word with stress on the fifth syllable — something like *милиционером* in Pasternak's *My Sister-Life* («Sestra moei zhizni»). In her "Poem of the End" Tsvetaeva has three verse lines in rhythmic form II; she uses a five-syllable notional word and, to meet the requirements of the meter, she replaces the first syllable with the conjunction *И*; thus, a six-syllable rhythmic word with a stress on the fifth syllable is formed:

И жмусь... И неотторжима n iS ‖ n i n n iS n
[Tsvetaeva 1965: 463]

■□

The *sixth* level of analysis involves using probability theory to analyze the statistics of rhythmic forms. The more frequently a word of a certain rhythmic type occurs in speech, the higher is the probability that such a word would meet the poet's needs at the necessary place. There are two ways of calculating the probabilities to correlate the possibilities offered by speech with the poetic text.

1) The frequencies of each rhythmic word are calculated on the basis of a suitably long and complex prose text.

What do the terms "suitably long" and "complex" imply? Tomashevsky used *Eugene Onegin*, though Kolmogorov did point out that a *prose* text should be used. A.V. Prokhorov analyzed Pushkin's "The Queen of Spades" ("Pikovaia dama"). Gasparov made his selections from Russian prose of the first half of the twentieth century. When I was examining versification of the 1950–1960s, I took random samplings from notable works by our prose writers of the sixties. There were also other experiments. To calculate the language probabilities of words properly, I consider it important to use prose texts from approximately the same period to which the poetic texts being studied belong.

Kolmogorov writes that, on the one hand, this factor may not significantly affect results of the research; on the other hand, the numbers might differ by 10 –20%, so that one would have to use them with care (Kolmogorov 1968: 152).

Then one calculates the probability that the words, suitable for the particular rhythmic form, come from the language – "without an effort on the poet's part." For this purpose the theorem of multiplying probabilities is used; to find the probability one must multiply the frequencies of the words in the given rhythmic form. The result obtained is compared with the frequencies of the rhythmic forms in a poetic text.

For instance, say we need to find out the theoretical frequency for rhythmic form I in the meter used by Tsvetaeva that we have been analyzing. Having selected a fairly long and complicated prose text, we calculate the frequency of the rhythmic words that comprise the form. For the first variant of that form, these would be a two-syllable word stressed on the second syllable, a three-syllable word with stress on the third syllable, and a two-syllable word with stress on the second syllable. We multiply the frequencies of these rhythmic words. After that, in the same way, we calculate the frequency for each of the other five variants of rhythmic form I, and multiply the frequencies of the rhythmic words which constitute each of the variants. Then we add up the frequencies for all six variants. Now we know the theoretical probability of this rhythmic form's occurrence in the text of a poem, "provided the poet does not interfere." We may then calculate the actual frequency of form I in Tsvetaeva's text and find out if she shows a preference for this form, avoids it, or employs the form with a frequency equal to its chance occurrence — i.e., is indifferent to it.

2) Fragments with the number and arrangement of syllables and stresses that correspond to different rhythmic forms are extracted from a prose text of adequate length. The result is compared with the frequency of rhythmic forms in the poetic text.

Kolmogorov expresses the results of his calculations as percentages, which, of course, does not change the matter. His data for the "Desk" cycle are as follows:

I	98	76.6%	43.8%
II	—	0.0	3.5
III	14	10.9	45.6
V	11	8.6	2.9
VI	1	0.8	1.3
VII	4	3.1	3.0
	128	100.0	100.1

The first column indicates the rhythmic forms; the second, the number of lines with that form in poems 1–5 of the cycle; the third shows that result as a percentage of the total number of lines; and the fourth, the theoretical percentage for the given rhythmic form.

In comparing the last two columns we see that Tsvetaeva employs rhythmic form I more than one and a half times the predicted amount, thereby overcoming the resistance of the language. At the same time she employs rhythmic form III four times less than the language would indicate, overcoming its resistance again. According to calculations of the theoretical percentages, rhythmic forms I and III have approximately the same probability of occurring in a text, considering only the norms of the language and not the poet's own role. In Tsvetaeva's texts we observe a sharp increase and decrease of those frequencies. Lines in which all the ictuses are stressed and all the non-ictuses are unstressed noticeably predominate.

■□

The *seventh* level of Kolmogorov's research is the statistics for word-boundaries. We shall only mention it here. Beginning with A. Bely, leading researchers of verse have been examining word-boundaries, with everyone absolutely carried away by this topic and everyone contradicting each other. I have not devoted much study to word-boundaries. They do need to be kept in mind when examining rhythmic forms and

calculating their actual significance against the data in the probability theory model, and we have just carried out such an analysis with the reader. But examination of word-boundaries by itself does not enrich our understanding of verse.

The *eighth* level of verse research, according to Kolmogorov, is examining the placement of different rhythmic form lines within the text; the *ninth* level is the study of enjambment and pauses between lines and entire stanzas. At the eighth level research is carried out intuitively, no specific methodology has been proposed. Over the forty years since the publication of this article, the study of enjambment (the ninth level), including that in Tsvetaeva's works, has made great advances. (For a very recent work on the subject, see Baevskii, Novikova, Romanova 2009).

But the methodology for examining rhythmic verse forms and their variants against the background of probability models for meters does not lag behind. "The Russian method" remains, in the words of Jakobson, an instance of an extremely long and remarkable alliance between linguistic poetics and the mathematical analysis of random processes (Jakobson 1961: 252).

■□

3. Support for exact methods in literary theory

In the 1960s–1980s Kolmogorov was at the height of his fame. He was opening new directions in various fields of mathematics and mechanics as well as in numerous realms of science more or less closely related to mathematics; he was establishing new scientific institutes, departments, and scientific journals; and he was being highly respected and honored in all these spheres of culture. Somewhere around the end of the 60s or beginning of the 70s, in a modest provincial department of literature where I was working, a colleague said to me: "But for the support you received from academician Kolmogorov, we would have wiped out all of you!" I was shocked by the hatred bubbling in his voice. M. L. Gasparov had to leave the Academy of Sciences' Institute of World Literature for the V.V. Vinogradov Russian Language Institute. As Gasparov told me, the reason was the unbearable atmosphere at the Institute of World Literature.

At Moscow State University Kolmogorov organized a small research group that became involved in studying verse theory by means of mathematical statistics and probability theory. As noted above, beginning in 1962 Kolomogorov's own articles began to appear in print, as did works he co-authored with his students and the students' independent papers. He also established a journal, *Teoriia veroiatnostei i ee primeneniia* (*Probability Theory and its Applications*), and in one of the first issues he published an article by M. L. Gasparov. Each of Kolmogorov's public presentations on verse theory would turn into a scholarly event, as is vividly described in Uspenskii (1997).

In 1962, after 11 years of working as a schoolteacher at a miners' settlement in the Donetsk region, I started to work at Smolensk Pedagogical Institute (which is now a university). I had to choose the main direction for my research. As a university student I had studied Turgenev's shift from the poetics of the natural school to the genre of the novel, and I even wrote a dissertation on this topic. However, by the time I started to work at the institute the dissertation had completely lost its attraction for me, so I decided not to defend it. Though my former fellow-student, the poet Felix Krivin, once said to me: "You should defend your dissertation; after that your dissertation will be defending you". But instead of following his advice I began looking through all the philological journals that were available at our regional academic library and at that of

the institute. In those days everyone was extremely impressed with Solzhenitsyn's "One Day in the Life of Ivan Denisovich" in *Novyi mir*. It goes without saying, that I did not avoid the general frenzy. And, I suppose, I was equally impressed with the article by Kolmogorov and A.M. Kondratov (1962) in *Issues of Linguistics* (*Voprosy iazykoznaniia*). This was Kolmogorov's first publication devoted to the study of verse theory. It dawned on me that this was precisely my field of research, and the very evening that I read the article I outlined the plan for a work that would develop one of Kolmogorov's ideas.

Memoirs about Kolmogorov, both those published and my own, indicate that he paid great attention to his correspondence and would respond to a stranger's letter almost immediately. The successes of verse studies during the sixties inspired me to consolidate them by publishing a collection of works that would emphatically mark those achievements. I invited a colleague, with whom I had become close when studying verse theory, to collaborate with me. We outlined a plan for the volume, and I took it upon myself to ask Kolmogorov's permission to republish two of his articles.

I wrote to him in June 1971, and got ready to wait for his answer. But it turned out I did not have to wait. The response came practically with the return post, dated 17June. Here is the entire letter, which I have never previously published. Among other things, in this letter Kolmogorov evaluates three of his works on verse theory.

June 17, 1971

Dear Vadim Solomonovich:

I am glad that our article "On the Bases of Russian Classical Metrics" has interested you. But reprinting it is not advisable, because we are going to change slightly the final formula for the schematic notation of rhythmic patterns and will soon publish it in revised form. Besides, I would like you to include in your collection those of my works in which the analysis of rhythm reveals the semantic significance of rhythmic changes. It is also important to illustrate the possibility of a statistical approach to studying texts that are relatively brief. Both aspects of this problem are rather thoroughly treated in the article "On the Meter of Pushkin's 'Songs of the Western Slavs'" (*Russkaia literatura*, 1966, № 1). "A Sample Study of a Meter and its Rhythmic Variants," which you suggested, is also suitable. I would not dare suggest both of these articles for your volume, but would like to remark that I consider the first one to be especially successful.

I would be happy to learn about what you are doing in regard to "more precisely defining the degree of syllable prominence within the verse line."

Yours, A. Kolmogorov

Unfortunately, we did not manage to publish that collection of articles. It turned out that the colleague whom I had invited to collaborate had a different view of the situation in the fireld of verse study: he was afraid to be attacked by demagogues for his support of probability theory and statistical models in verse theory and simply did not how to use them. On my own I failed not only to publish but even to collect the articles. Some years later, after I had defended my dissertation on the "Verse Typology of Russian Lyrics," which was entirely based on statistical models, that man asked me to take a look at his doctoral dissertation, which he was about to defend. I did what he asked and then told him that I was very surprised by his characterization of David Samoilov's verse as that of a traditionalist. I also told him that the statistical method I

had been using proved Samoilov's verse to be innovative, but when I was just about to show him the exact figures he stopped listening to me, took out a pen, and simply crossed off Samoilov's name from the column of traditional poets and put it into the column of innovators.

When I sent my work to Kolmogorov in response to his request, he again answered almost immediately (24 July 1971). The letter began with an inspiring phrase:

"Dear Vadim Solomonovich, I consider the basic direction of your research to be interesting and necessary".

Long ago A. Bely already faced difficulties in calculating the number of stressed and unstressed syllables due to the great number of prepositions, conjunctions, particles and interjections, which some researchers considered to be stressed and others as unstressed. They were also called metrically ambiguous or semi-stressed. Some scholars also included in this category pronouns, adverbs, numerals and auxiliary verbs (Belyi 1929: 239–247; Zhirmunskii 1925: 126–130; Piast 1931. 332–345). Kolmogorov too faced these difficulties. In the work I had sent him I suggested overcoming these obstacles with the help of the "Potebnya effect." In one of his articles, which was buried in a provincial scholarly journal from the middle of the nineteenth century, this remarkable linguist, who liked to be published in such unpretentious outlets, suggested determining syllabic prominence in Russian speech by means of a special system that is more subtle than the dichotomy "stressed" / "unstressed" (Potebnia 1865: 62). I decided to use his method to eliminate the problem of "semi-stressed" words in poetic speech. However, Kolmogorov rejected my suggestion: his own assessments concerning the prominence of certain syllables did not coincide with those of the "Potebya effect".

Nevertheless, Kolmogorov gave me some valuable advice in case I was going to continue using the "Potebnya effect" in my research. I thought this meant that he did not completely reject my attempts. I decided not to argue with this great person and only thanked him kindly.

At that time it did not even occur to me that my study of verse theory could grow into a doctoral dissertation, for I had become absorbed in this remarkable field out of sheer interest. And furthermore, inveterate demagogues were declaring that research on versification, especially when employing mathematical statistics and probability theory, was pure formalism. And formalists were regarded virtually as public enemies, a fifth column in literary theory.

About five years ago, when I began collecting the material for articles and lectures about Kolmogorov as a scholar of verse theory, I came across this aphorism of his: "You should be indifferent to your dissertation" [1, p. 27]. He would say that devotion to scholarship would lead naturally to clear success in research. And this was exactly what happened to me, a follower of Kolmogorov. I was not engaged in any human rights movement during the sixties and seventies, though I hated the false totalitarian regime that ruled our country and that finally collapsed in the eighties under the weight of its lies and transgressions. I expressed my protest through my scholarly and literary work, as well as through teaching.

It was not a rare thing for Kolmogorov not only to meet with enthusiastic approval, but also to face hostility and total lack of understanding, even among mathematicians. Once one of his graduate students was flunked when defending his candidates' dissertation. Anticipating possible attacks by incompetent ignoramuses, he found it necessary to begin one of his classical papers on probability theory with a very simple explanation of the most basic concepts: "If one intends to examine natural or

social phenomena by means of mathematics, these phenomena need to be schematized beforehand. The point is that, when investigating the process of change in a given system, mathematical analysis can be applied only if any possible state of this system can be fully determined with the help of a known mathematical apparatus — for example, the values of a certain number of parameters. Such a mathematically determinable system is not reality itself, but a scheme suitable for the description of reality" (Kolmogorov 1986: 61). The work quoted here was first published in German in 1931; a Russian translation came out in 1938.

Kolmogorov's works on verse theory were met with similar attacks in literature journals. An ignorant critic of "probability and statistical verse study" reproaches him for studying not verse itself but its schematic representation. "The result of his method is two isolated groupings: the actual verse material and a hypothetical, abstract closed system, some kind of an 'ideal model,' unacceptably remote from verse and having its own immanent laws based on probability theory" (Goncharov 1973: 16)..

The year 1975 was a remarkable one in my life. At the very beginning of that year, thanks to the support of Iurii Lotman, I defended in Tartu my doctoral dissertation, "Verse Typology of Russian Lyrics," and toward the end of 1975 *Izvestiia AN SSSR, seriia literatury i iazyka* published my article, written in collaboration with P.A. Rudnev, defending Kolmogorov and A.V. Prokhorov from ignorant demagogues. Prokhorov was not only Kolmogorov's student and co-author, but also a person very close to him. We wrote: "The rhythm of phonological units is a specific feature of poetic speech, distinguishing it from prose and transforming other features and levels. Abstracting and studying this component of the system are extremely important tasks, which are being worked on particularly by those scholars against whom the controversy is aimed – academician Kolmogorov and A.V. Prokhorov. No study of verse rhythm by these scholars says or implies that the characteristics of poetic speech are limited to the properties of rhythm; on the contrary, in their most successful papers rhythm is studied in correlation with the lexicon, semantics and phonetics. As for probability theory, it is not considered to be the basis for rhythmic norms – it is solely a means for their analysis. It is necessary to note that in the works on versification by academician Kolmogorov the application of probability methods is strictly limited; such an application is always preceded by a subtle philological analysis. The precepts about meter by academician Kolmogorov — which are familiar only through individual observations scattered among his numerous articles, in a far from complete form, and which cannot be reduced to the several popular quotations that have 'wandered' from one article to another — are extremely substantial. They are correlated with fundamental tenets (a) of modern linguistics, regarding the unity between the level of content and that of its expression; (b) of literary theory, regarding the author's role in the literary process; and (c) of dialectical aesthetics, regarding the unity of form and content in a work of art" (Baevskii, Rudnev 1975: 441).

Our article in collaboration with Rudnev caused a real storm. Even some of those who understood the important role of Kolmogorov and therefore supported our views, were displeased: they feared repressions against the entire movement. Looking back at that time, I clearly see that no harm was done by our article. It actually helped us discharge the obligation of each researcher to uphold his scholarly ideas and contributed a bit to the radical change of attitude toward the mathematical study of verse theory that occurred during the 1980s.

My doctoral dissertation happened to be the first one devoted to verse theory in the USSR. And as far as I know, it was also the first doctoral dissertation in a humanities discipline that was based entirely on mathematical statistics and probability theory, with its supplements covering 490 pages in all. It was natural for VAK[5] to request the most competent Soviet specialist in the field of literary theory to provide a peer review of my work. His review was devastating. And then VAK took a remarkable step: the dissertation was re-directed to academician Kolmogorov for a second review — about which I was not aware.

On 1 September 1976, I came to my place of after summer vacation. The secretary at the dean's office gave me a letter from Kolmogorov, dated 1 August 1976. In this letter he informed me that VAK had sent him my dissertation and that a copy of his peer review was enclosed. He also inquired whether I could send him my works devoted to the subject of my dissertation and, for a while, the dissertation itself. In those days there were no personal computers. Dissertations would be prepared on a typewriter using carbon paper. This method allowed for no more than five copies. One copy was requested by Iu. M. Lotman, just after the defense (he was my opponent). Another copy was expected to be given to the Tartu University Library. The third copy I owed to the dissertation room of the former Lenin Library in Moscow. The fourth copy I happily sent to Kolmogorov as a present. As for the fifth, I kept it for myself, and at this very minute it lies before me. The edition, I believe, had quite a good distribution.

When studying versification, I examined not only its immanent features but its links with poetic speech as well. I did not analyze the works of hundreds of poets but selected only the twenty authors from the most recent poetic epoch of 1956–1965 who received the greatest critical attention—ranging from Akhmatova to Akhmadulina. I randomly selected one thousand poems, and for each text examined twenty variables related to the features of meter, rhythm, stanzaic form, rhyme, poetic syntax, thematics and the system of imagery.

After that, with the help of several simple statistical methods, I compared each of these features with the other six. Sometimes variables within one and the same feature were compared.

Anna Lukianovna Belianovskaia, my wonderful teacher in grades one through four, taught me that three apples can not be added to four pears and two plums. But you can find the sum of three, four and two fruits. I had to search for the method to help me compare, for instance, the variables of verse meter and themes, or stanza organization and the imagery system. I was eager to know if the metrical structure predetermines, at least to some extent, the organization of themes or the structure of the imagery. Or is each of these features completely independent? And what about the variables within each feature?

And then I made a discovery, the most difficult and the most fruitful for the entire dissertation. Shortly before I started working on my dissertation I had been fond of the books by Ashby, one of the forefathers of cybernetics. He attaches great importance to the category of diversity, which is closely related to the concept of information and is easy to quantify (Ashby 1959: 173–274; Ursul 1968: 59–78). I calculated the diversity of each variable and then compared the coefficients of diversity among themselves. In rare instances I obtained a strong positive correlation, more

[5] The abbreviation 'VAK' stands for 'Higher Attestation Commission' – the authority in the former USSR and in present-day Russia that has been in charge of the entire system for awarding advanced academic degrees and academic ranks. *(Translator's note)*

frequently there was no correlation between the variables at all, and the most frequent case was when the correlation was positive but weak. In general, I was able to state that web-like links among variables and features prevail in a verse system and that the components of a system predominantly show a weakly positive correlation.

This result fully coincided with the notions of Kolmogorov and Lotman, which they had expressed intuitively. Guided by their opinion, I carefully extrapolated my results to artistic systems in general.

Kolmogorov's large and substantial peer review is dated 29 October 1976. Let us quote a characteristic excerpt (the emphases are by Kolmogorov himself).

"The statistical method plays an essential role in this dissertation. There has never been a statistical investigation in versification carried out on a large body of material and dealing with such a great number of characteristics. The author is successful in his broad application of a rank correlation between characteristics.

The mathematical statistics employed by the author are elementary. Nevertheless, many results of the statistical analysis are supported by his conceptual interpretation and are of significant value.

Valid statistical results are supplemented with general conclusions, which are supported by carefully selected examples that generally show the author's good taste. And on the whole, the author's remarkable erudition – the 553 works listed in the bibliography are utilized with great competence – makes reading the dissertation extremely interesting".

VAK sent me the peer review of the first reviewer and that of Kolmogorov, and summoned me to the presidium to re-defend my dissertation. I had no idea how it would turn out. In fact, what actually did happen cannot be called a real defense. I entered a cramped room where there were about fifteen complete strangers, some middle-aged and others rather old. I immediately sensed a friendly atmosphere. I was met with smiles. Practically at that very moment the first congratulations began. Kolmogorov's evaluation had a magic effect on the Higher Attestation Commission.

Meanwhile, at the Gorky Institute of World Literature M.L. Gasparov was preparing to defend his doctoral dissertation. Some obstacles would turn up, though I cannot say what they were exactly: Gasparov said little about them, and now I do not remember much of it. Everything settled down when the Dissertation Council allowed Kolmogorov to be the official opponent. A struggle over that had taken place. The fact that Gasparov was awarded his doctorate only in 1978 proves how tense that struggle was.

After that, defenses of doctoral dissertations on verse theory no longer met with serious obstacles. Another two or three were defended before 'perestroika' started—i.e., by the middle of the 1980s, when Kolmogorov was still alive. He paved the way for this direction in scholarship. Since then, dissertation defenses on verse theory have become as non-dramatic as those of other philological and mathematical dissertations.

References

[1] *Kolmogorov v vospominaniiakh uchenikov.* [Kolmogorov in the Recollections of his Students]. (2006). Moscow: MTsNMO.

[2] *Kolmogorov: Iubileinoe izdanie v trekh knigakh.* [Kolmogorov: Anniversary Issue in Three Volumes]. (2003). Moscow: Fizmatlit, Vol. 2.

Ashby, U.R. (1959). *Vvedeniie v kibernetiku.* [Introduction to Cybernetics]. Moscow: Inostrannaia literatura.

Baevskii, V.S. (1976). *Ritmicheskaia kompozitsiia stikhotvoreniia. In: Zhanr i kompozitsiia literaturnogo proizvedeniia.* [The Rhythmic Composition of a Poem. In: Genre and Composition of a Literary Work]. Kaliningrad.

Baevskii, V.S. (1982). *Uchebnyi material po analizu poeticheskikh tekstov.* [Study Material for the Analysis of Poetic Texts]. Tallinn.

Baevskii, V.S. (2003). *Istoriia russkoi literatury XX veka.* [The History of Twentieth-Century Russian Literature]. Moscow: Iazyki slavianskikh kul'tur.

Baevskii, V.S., Novikova, M.S., Romanova, I.V. (2009). Sintaksicheskii perenos (enjambment) ot Shekspira do Brodskogo: ontologicheskii argument i statisticheskie dannye. [Enjambement from Shakespeare to Brodskii: an Ontological Argument and Statistical Data]. In: *Slavianskii stikh VIII: Stikh, iazyk, smysl: 218-232*. Moscow: Iazyki slavianskoi kul'tury.

Baevskii, V.S., Rudnev, P.A. (1975) Stikhorusistika – 73. [Russian Verse – 73]. *Izvestiia AN SSSR, seriia literatury i iazyka, № 5.*

Bagritskii, E. (1934). *Odnotomnik*. [Omnibus Volume]. Moscow: Sovetskaia literatura.

Belyi, A. (1929). *Ritm kak dialektika i "Mednyi Vsadnik"*. [Rhythm as Dialectics and "The Bronze Horseman"]. Moscow: Federatsiia.

Blok, A.A. (1960). *Sobranie sochinenii v vos'mi tomakh* [Collected Works in 8 volumes*], Vol. 3.* Moscow – Leningrad: GIKhL.

Blok, A.A. (1963). *Sobranie sochinenii v vos'mi tomakh* [Collected Works in 8 volumes], *Vol.7*. Moscow – Leningrad: GIKhL.

Goncharov, B.P. (1973). O strukturalizme v stikhovedenii. [On Structuralism in Verse Study]. *Filologicheskie nauki, No. 1, 3-17.*

Jakobson, R. (ed.) (1961). *Structure of Language and its Mathematical Aspects*. Providence.

Kolmogorov, A. N. (1963). K izucheniiu ritmiki Maiakovskogo. [On the Study of Maiakovskii's Rhythmic System]. *Voprosy iazykoznaniia. 4, 84-95*. Moscow: Izdatel'stvo AN SSSR.

Kolmogorov, A.N. (1965). Zamechaniia po povodu analiza ritma "Stikhov o sovetskom passporte Maiakovskogo". [Comments on the Rhythmic Analysis of Maiakovskii's "Verses about a Soviet Passport"]. *Voprosy iazykoznaniia. 3, 70-75*.. Moscow: Nauka..

Kolmogorov, A.N. (1966). O metre pushkinskikh "Pesen zapadnykh slavian [On the Meter of Pushkin's "Songs of the Western Slavs"]. *Russkaia literatura No. 1, 98-111.*

Kolmogorov, A.N. (1968). *Primer izucheniia metra i ego ritmicheskikh variantov*. In: *Teoriia stikha*: *145–67.* [A Sample Study of a Meter and its Rhythmic Variations. In: Verse Theory]. Leningrad: Nauka,

Kolmogorov, A.N. (1984). Analiz ritmicheskoi struktury stikhotvoreniia A.S. Pushkina "Arion". In: *Problemy teorii stikha. 118-124*. [An Analysis of the Rhythmic Structure of A.S. Pushkin's "Arion". In: The Problems of Verse Theory]. Leningrad: Nauka.

Kolmogorov, A.N. (1986). Ob analiticheskikh metodakh v teorii veroiatnostei. [On an Analytical Approach in Probability Theory] In: Kolmogorov A. N., *Probability Theory and Mathematical Statistics*. Moscow: Nauka.

Kolmogorov, A. N., Kondratov, A. M. (1962). Ritmika poem Maiakovskogo. [The Rhythmic System of Maiakovskii's Narrative Poems]. *Voprosy iazykoznaniia 3* 62-74. Moscow: Izdatel'stvo AN SSSR

Kolmogorov, A.N., Prokhorov, A.V. (1963). O dol'nike sovremennoi russkoi poezii (obshchaia kharakteristika). [The Dol'nik in Modern Russian Poetry (General Characteristics)]. *Voprosy iazykoznaniia. 6, 84-95.* Moscow: Izdatel'stvo AN SSSR.

Kolmogorov, A. N., Prokhorov, A.V. (1964). O dol'nike sovremennoi russkoi poezii (statisticheskaia kharakteristika dol'nika Maiakovskogo, Bagritskogo, Akhmatovoi). [The Dol'nik in Modern Russian Poetry (Statistical Characteristics of the Dol'nik in the Poetry of Maiakovskii, Bagritskii and Akhmatova)]. *Voprosy iazykoznaniia 1, 75-94.*. Moscow: Nauka.

Kolmogorov, A.N., Prokhorov, A.V. (1968). K osnovam russkoi klassicheskoi metriki. In: *Sodruzhestvo nauk i tainy tvorchestva. 397-432.* [On the Bases of Russian Classical Metrics. In: Collaboration of Sciences and the Mysteries of Art]. Moscow: Iskusstvo.

Kolmogorov, A.N., Prokhorov, A.V. (1985). Model' ritmicheskogo stroieniia ruskoi rechi, prisposoblennaia k izucheniiu metriki klassicheskogo stikha. In: *Russkoe stikhoslozhenie: 113-134.*. [A Model of Rhythmic Patterns in Russian Speech Applied to Analyzing the Metrics of Classical Verse // Russian Versification]. Moscow: Nauka

Kolmogorov, A.N., Rychkova, N.G. (1999). Analiz ritma russkogo stikha i teoriia veroiatnostei. [Analyzing the Rhythm of Russian Verse and Probability Theory] *Teoria veroiatnostei i ee primeneniia 44(2), 19 – 431.*

Krechmer, E. (1927). *Meditsinskaia psikhologiia*. [Medical Psychology]. Moscow.

Piast, V. (1931). *Sovremennoe stikhovedenie. Ritmika.* [Modern Verse Study. Rhythmics]. Leningrad: Izdat. pisatelei v Leningrade.

Potebnia, A.A. (1865). *O zvukovykh osobennostiakh russkikh narechii.* [On the Phonic Distinctive Features of Russian Adverbs] In: *Philologicheskiie zapiski, Iss.1.*

Samoilov, D. (2006). *Stikhotvoreniia*. [Poems]. St.Petersburg: Academicheskii proiekt.

Taranovskii, K. (1953). *Ruski dvodelny ritmovi, I-II.* [Russian Binary Meters]. Belgrade: Srpska akademija nauka.

Tomashevskii, B.V. (1971). Pis'ma V. Ia. Briusovu. In: *Trudy po znakovym sistemam 5.* [Letters to V. Ia. Briusov. In: Works on Semiotic Systems]. Tartu: Tartu University Press.

Tomashevskii, B.V. (1929*). O stikhe*. [On Verse]. Leningrad: Priboi.

Tsvetaeva, M.I. (1965). *Izbrannye proizvedeniia.* [Selected Works]. Moscow – Leningrad: Sovetskii pisatel'.

Ursul, A.D. (1968). *Priroda informatsii*. [The Nature of Information]. Moscow: Politisdat.

Uspenskii, V.A. (1997). Kolmogorov i semiotika. [Kolmogorov and Semiotics] *Novoe literaturnoe obozrenie 24.*

Zhirmunskii, V. M. (1925). *Vvedeniie v metriku. Teoriia stikha.* [Introduction to Metrics. Verse Theory]. Leningrad: Academia.

Translated by N. V. Kutuzova

Glottometrics 22, 2011, 44-56

Word length in Slovak poetry

Radek Čech[1]
Ioan-Iovitz Popescu
Gabriel Altmann

Abstract. In the present article the place of word-length in Slovak poetry in the framework of the general theory is sought. Different models are presented. The possibility of applying Menzerath's law in this domain is scrutinized.

Keywords: word length, verse length, Slovak, binomial distribution, Poisson distribution, Ord's scheme, Menzerath's law

1. Introduction

Word-length is a problem which has occupied generations of both linguists and mathematicians. The literature is enormous (cf. Best 2001, 2006; Grzybek 2006) but the number of problems associated with word length seems to increase, e.g., the distribution of word length in different languages, word properties associated with word length, etc. The latter problem belongs to the domain of synergetic linguistics where word length plays a central role already since G.K. Zipf (1935/1968) (cf. Köhler 2005).

The first problem has two aspects: (1) Word length is a general phenomenon obeying some laws (except for monosyllabic languages where it is not a variable), and (2) even within one and the same language it is characteristic for text sorts, style, author, etc. (cf. Best 2001, 2006; Grzybek 2006) In both cases boundary conditions play an enormous role, but if a model functions in about 90% of cases, one usually does not care any more for subsuming the rest of the cases under the given law or searching for the causes of deviation. Usually, 5% of bad fits is no reason for a rejection. However, some authors do it, introduce modifications of the model or derive the distribution of word length in a quite different way (e.g. Uhlířová 1995; Wimmer, Witkovský, Altmann 1999). Though the history abounds in models (cf. Grzybek 2006a), a theory be cannot easily constructed. Language must fulfil a number of requirements. These are not the general requirements listed by Köhler (2005) but specific ones, so to speak local ones enabling the author/writer to write a text for exactly the given occasion. He must express something, express himself and give the text an adequate form. Seen from this perspective, every text is a unique creation, hence mixing texts, e.g. taking a corpus as a whole and scrutinise in it the word-length distribution, may lead to distortions.

A well known problem connected with testing a model is the character of the sample. Sometimes the number of different length classes is too small for testing a model, e.g. the words are not longer than four syllables and the model has three para-

[1] Radek Čech, Department of Czech Language, University of Ostrava, Reální 5, Ostrava, 701 03, Czech Republic, e-mail: radek.cech@osu.cz,

meters. The chi-square test for goodness-of-fit cannot be applied because there are no degrees of freedom. Some other tests for fitting are obsolete and should not be used. The chi-square itself is not quite safe because the result depends on the sample size (cf. Rietveld et al. 2004). Thus the creation of a theory has both theoretical and empirical hindrances.

In our analysis we shall try to characterize Bachletová's poetry as a solid block of length distributions placed in a restricted area (Chapter 2). The distributions are always binomial being a further characteristic of the author (Chapter 3). We further try to find a distribution of verse length in terms of word numbers (Chapter 4) and lastly the relation of mean word length to the verse length, a view that has been neglected so far.

2. Methodology

In the present article we restrict ourselves to one author, Eva Bachletová, and her poetic texts written in Slovak. The poems by Bachletová are rhymeless, not following a metric prescription, and the individual verses are rather short, many times consisting only of one word.

We count for each poem separately the number of words having length x = 1,2,3,… measured in terms of syllable numbers. The length zero which is frequent in Slavic languages has been omitted because the pertinent words are consonantal prepositions (e.g. Slovak *k, s, v, z*) which can be considered proclitics of the next word. Their separate counting leads to a modification of every model (cf. Uhlířová 1995; Wilson 2006). Hence, the numbers in the second column of Table 1 are to be read as follows: in the poem *Aby spriesvitnela* there are 14 words of length 1; 26 words of length 2; 15 words of length 3; 5 words of length 4, and 4 words of length 5. All operations are performed upon these empirical data, e.g. the average length of verses is [1(14) + 2(26) + 3(15) + 4(5) + 5(4)]/64 = 151/64 = 2,3594 words.

An empirical distribution can be characterized by many means, e.g. moments and their functions (asymmetry and excess), entropy, repeat rate, etc. Here we shall use two functions of moments proposed by J.K. Ord (1972) used frequently in text analysis, namely

$$(1) \qquad I = \frac{m_2}{m_1'}$$

and

$$(2) \qquad S = \frac{m_3}{m_2},$$

where $m_1' = \bar{x}$, the mean of the distribution, and m_2 and m_3 are the second and third central moments defined as

$$m_r = \frac{1}{N}\sum_{x=1}^{k}(x - m_1')^r f(x) \quad (3)$$

where N is the sum of the frequencies $f(x)$ and k is the greatest length. The second central moment (r = 2) is the variance (i.e. a measure of dispersion) and the third moment (r = 3) is the indicator of skewness of the distribution. Hence, I and S are functions of some properties of the distribution.

The results of counting are presented in Table 1. The frequencies in the column "Distribution" represent those of lengths x = 1,2,3,…

Table 1
Word-length distributions and Ord's criterion <*I*,*S*> in poems by E. Bachletová

Poem	Distribution	I	S
Aby spriesvitela	14,26,15,5,4	0,5082	0,8249
Bez rozlúčky	15,13,5	0,3030	0,3739
Čakáme šťastie	8,19,9,6,2	0,4747	0,6625
Čakanie na boží jas	30,24,18,2	0,3938	0,3986
Čas pre nádych vône	23,36,16,4,0,1	0,4284	1,0712
Dielo Stvoriteľa	41,53,23,13,1	0,4545	0,6582
Dnešný luxus	14,9,8,3,1	0,5855	0,7951
Do večnosti beží čas	15,22,11,1	0,3116	0,2563
Hľadanie odpovedí	22,20,18,4	0,4223	0,3145
Iba neha	51,53,16,5,0,2	0,4925	1,5436
Iba život	10,18,10,5	0,3924	0,3497
Idem za Tebou	26,28,10,5	0,4203	0,6943
Ihly na nebi	26,18,7,1	0,3614	0,6887
Keď dohorí deň	22,16,11,2	0,4215	0,5399
Kým ich máme	16,20,5,1,1	0,4145	1,1458
Len áno	7,17,5,1	0,2867	0,3750
Malé modlitby	13,25,8,2	0,3050	0,4421
Malý ošiaľ	32,22,12,1	0,3721	0,5581
Mladé oči	7,8,3,1	0,3830	0,6075
Moje určenie	55,54,25,6,2	0,4448	0,8493
Neopusť ma	6,17,7,1,0,1	0,4432	1,6529
Náš chrám	28,26,19,6,1,1	0,5509	0,9998
Naše dejiny	6,7,6,5,1	0,5435	0,2941
Naše mamy	22,19,11,4	0,4481	0,5909
Naše svetlo	16,23,11,7	0,4356	0,4740
Neha domova	10,11,3,1	0,3556	0,6750
Nepoznateľné	39,33,12,4,1,1	0,5381	1,4643
Podobnosť bytia	23,32,22,5,1,1	0,4726	0,9170

Prvotný sen	23,30,13,9,2	0,5206	0,7757
Rozdelená bytosť	26,31,16,3	0,3634	0,4184
Rozťatá prítomnosť	30,34,9,3	0,3507	0,6667
Som iná	20,24,6,4,1,1	0,5928	1,5737
Spájania	18,14,8,2	0,4249	0,6133
Stály smútok pre šesť písmen	54,64,19,4	0,3287	0,5505
Tak málo úsmevu	19,25,11,5,1	0,4614	0,7605
Tiché verše	8,10,11	0,3064	-0,1515
To všetko je dar	16,18,12,1	0,3469	0,2531
Večerná ruža	13,18,9,2,1,1	0,5672	1,4309
Večerné ticho	25,27,9,5	0,4242	0,7226
Vo večnosti slobodná	38,71,44,5,0,1	0,3417	0,5724
Vrátili sa	17,18,9,4	0,4375	0,5714
Vyznania	17,25,9,3	0,3578	0,5331
Z neba do neba	13,27,18,5,0,1	0,4174	0,8585
Zasľúbenie jasu	12,23,10,3	0,3367	0,4026
Zbytočné srdce	12,15,5,4	0,4517	0,6749

If we plot <I,S> in a Cartesian coordinate system, we obtain the results presented in Figure 1. One can see that the points are placed on a straight line with relative great dispersion. The trend can be expressed by S = -0,5842 + 3,0397I yielding an R^2 = 0.41 which is small, but in poetry of this kind – without any binding – it is sufficient. As a matter of fact, continuing to evaluate more poems of the author one would obtain an ellipse, but preliminarily we only want to show the unity of the author. The ellipse can be constructed using our results: the slope of the longer axis is given by the regression coefficient of the straight line and the shorter axis is given as $1 - R^2$ placed orthogonally to the mean of the long axis. The straight line $S = 2I - 1$ represents the upper boundary of the beta-binomial (negative hypergeometric) distribution and serves here for orientation.

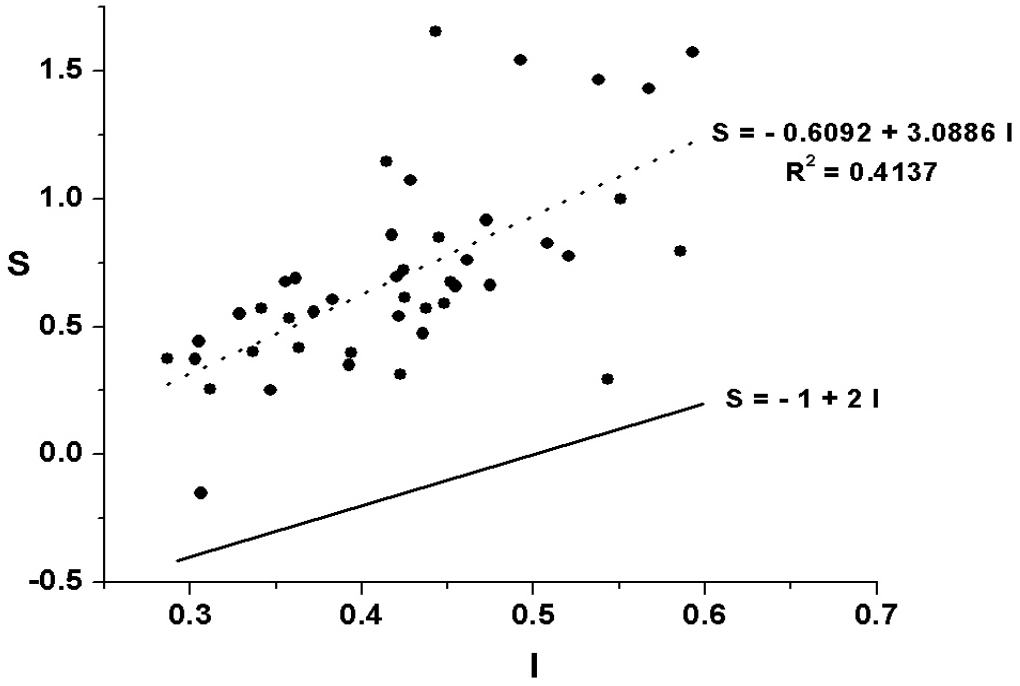

Figure 1. The <I,S> domain of word-length distributions with Bachletová

The model of the individual distributions is not easy to set up. First, the support of the data is very short. Though all data lie in the domain of Ord's hypergeometric distribution ($1 > S > 2I - 1$, $I < 1$) one needs at least five well represented length classes in the empirical distribution in order to have at least 1 degree of freedom. Evidently we must test models with smaller number of parameters. As candidates there are the Poisson and the binomial distributions which are limiting cases of the hypergeometric distribution. (Poisson when $N \to \infty$, $M \to \infty$, $n \to \infty$, $nM/N \to a$; binomial when $N \to \infty$, $M \to \infty$, $M/N \to p$). As a matter of fact, the great majority of data abide by these models. The points above $S > 1$ belong to the domain of the beta-Pascal distribution which has, unfortunately, three parameters and is not useful here. In Table 2, these distributions are fitted to the individual poems and tested using the chi-square test for goodness-of-fit. Of course, both distributions are displaced 1 step to the right because the support of the data is always x = 1,2,…,*n*. The formulas are

Binomial distribution: $P_x = \binom{n}{x-1} p^{x-1} q^{n-x}, \; x = 1,2,...n+1$

Poisson distribution: $P_x = \frac{a^{x-1} e^{-a}}{(x-1)!}, \quad x = 1,2,...$

The right truncated Poisson distribution would be more realistic because length data can never be infinite, but we did not obtain good results using it. Besides, it has one parameter more. This caused the impossibility to apply several times the binomial distribution, too.

Table 2
Fitting the Poisson and the binomial distributions to word-length data
in poems by E. Bachletová

Poem	**Distribution**	**Parameters**	**X^2**	**DF**	**P**
	Poisson	a = 1,3871	1,67	3	0,64
Aby spriesvitela	Binomial	n = 1374, p = 0,0010	1,67	2	0,43
Bez rozlúčky	Poisson	a = 0.7408	0,26	1	0,61
	Binomial	n = 7, p = 0,3182	2,60	3	0,46
Čakáme šťastie	Poisson	a = 2,2538	5,14	4	0,27
	Poisson	a = 0,9441	4,48	2	0,11
Čakanie na boží jas	Binomial	n = 6; p = 0,1529	3,70	1	0,05
	Binomial	n = 5; p = 0,2091	0,56	1	0,45
Čas pre nádych vône	Poison	a = 1,0588	3,09	3	0,38
	Poisson	a = 1,1161	3,90	3	0,27
Dielo Stvoriteľa	Binomial	n = 9; p = 0,1230	3,33	2	0,19
	Poisson	a = 1,0914	1,81	2	0,40
Dnešný luxus	Binomial	n = 1088; p = 0,0010	1,82	1	0,18
	Binomial	n = 3; p = 0,3216	0,30	1	0,59
Do večnosti beží čas	Poisson	a = 1,0206	4,00	2	0,14
	Poisson	a = 1,1189	3,43	2	0,18
Hľadanie odpovedí	Binomial	n = 6; p = 0,1799	2,82	1	0,09
	Poisson	a = 0,8616	2,17	3	0,54
Iba neha	Binomial	n = 864, p = 0,0010	2,16	2	0,34
	Poisson	a = 1,2761	0,95	2	0,62
Iba život	Binomial	n = 4; p = 0,3130	0,58	1	0,44
	Poisson	a = 0,9361	0,63	2	0,73
Idem za Tebou	Binomial	n = 925; p = 0,0010	0,62	1	0,43
	Poisson	a = 0,6852	0,40	2	0,82
Ihly na nebi	Binomial	n = 7; p = 0,0969	0,24	1	0,63
	Poisson	a = 0,8926	1,71	2	0,42
Keď dohorí deň	Binomial	n = 10; p = 0,0083	1,67	1	0,20
	Poisson	a = 0,8562	1,99	2	0,37
Kým ich máme	Binomial	n = 6; p = 0,1421	1,22	1	0,27
	Binomial	n = 3; p = 0,3337	1,84	1	0,18
Len áno	Poisson	a = 1,0219	5,56	2	0,06
	Binomial	n = 3; p = 0,3281	1,41	1	0,23
Malé modlitby	Poisson	a = 1,0018	5,25	2	0,07

	Poisson	a = 0,7653	2,27	2	0,32
Malý ošiaľ	Binomial	n = 5; p = 0,1493	2,01	1	0,16
	Poisson	a = 0,9120	0,26	2	0,88
Mladé oči	Binomial	n = 904; p = 0,0010	0,26	1	0,61
	Poisson	a = 0,9174	0,41	3	0,94
Moje určenie	Binomial	n = 13, p = ß,0707	0,28	2	0,87
	Binomial	n = 5; p = 0,2310	2,33	1	0,13
Neopusť ma	Poisson	a = 1,1701	4,80	2	0,09
	Poisson	a = 1,1173	0,91	3	0,82
Náš chrám	Binomial	n = 1123; p = 0,0010	0,91	2	0,63
	Poisson	a = 1,5794	1,53	3	0,68
Naše dejiny	Binomial	n = 1548; p = 0,0010	1,52	2	0,47
	Poisson	a = 0,9660	0,26	2	0,88
Naše mamy	Binomial	n = 958; p = 0,0010	0,26	1	0,61
	Poisson	a = 1,2015	0,51	2	0,77
Naše svetlo	Binomial	n = 1181; p = 0,0010	0,51	1	0,48
	Poisson	a = 0,8167	0,71	2	0,70
Neha domova	Binomial	n = 808; p = 0,0010	0,70	1	0,40
	Poisson	a = 0,8493	0,50	2	0,78
Nepoznateľné	Binomial	n = 859; p = 0,0010	0,50	1	0,48
	Binomial	n = 7; p = 0,1701	0,81	2	0,67
Podobnosť bytia	Poisson	a = 1,2003	2,05	3	0,56
	Poisson	a = 1,2049	1,93	3	0,59
Prvotný sen	Binomial	n = 1194; p = 0,0010	1,93	2	0,38
	Binomial	n= 4; p = 0,2375	0,14	1	0,70
Rozdelená bytosť	Poisson	a = 0,9810	2,31	2	0,31
	Poisson	a = 0,8171	2,55	2	0,28
Rozťatá prítomnosť	Binomial	n = 4, p = 0,2040	1,31	1	0,25
	Poisson	a = 0,8812	2,94	2	0,23
Som iná	Binomial	n = 1000; p = 0,0010	2,94	1	0,09
	Poisson	a = 0,8757	0,47	2	0,79
Spájania	Binomial	n = 868; p = 0,0010	0,47	1	0,47
Stály smútok pre	Binomial	n = 3; p = 0,2716	1,26	1	0,26
šesť písmen	Poisson	a = 0,8273	5,93	2	0,05
	Poisson	a = 1,0914	0,79	3	0,85
Tak málo úsmevu	Binomial	n = 1089, p = 0,0010	0,79	2	0,67
Tiché verše	Poisson	a = 1,3057	0,01	1	0,92
	Binomial	n = 3; p = 0,3221	1,17	1	0,28
To všetko je dar	Poisson	a = 1,0330	3,43	2	0,18
	Poisson	a = 1,0576	0,57	2	0,75
Večerná ruža	Binomial	n = 9; p = 0,1234	0,11	1	0,74
	Poisson	a = 0,9329	0,87	2	0,65
Večerné ticho	Binomial	n = 922; p = 0,0010	0,87	1	0,35
Vo večnosti slobodná	Binomial	n = 5; p = 0,2287	6,12	2	0,05

	Poisson	a = 1,0244	0,01	2	0,99
Vrátili sa	Binomial	n = 1015; p = 0,0010	0,01	1	0,92
	Binomial	n = 4; p = 0,2432	0,69	1	0,41
Vyznania	Poisson	a = 0,9809	2,22	2	0,33
	Binomial	n = 5; p = 0,2555	0,58	2	0,78
Z neba do neba	Poisson	a = 1,3094	4,10	3	0,25
	Binomial	n = 3; p = 0,3568	0,69	1	0,41
Zasľúbenie jasu	Poisson	a = 1,0857	2,82	2	0,24
	Poisson	a = 1,0746	0,97	2	0,61
Zbytočné srdce	Binomial	n = 1051; p = 0,0010	0,98	1	0,32

The results of fitting are very persuading. There is no exception; all fittings are significant. In some cases only the Poisson distribution was applicable, because the number of classes was too small for the binomial (*Tiché verše; Bez rozlúčky*); in one case only the binomial was applicable (*Vo večnosti slobodná*). In many cases one can see that the binomial distribution converges towards the Poisson: this is evident in cases where the parameter n is very great and p is very small (usually 0,0010 because of computing restriction). The product np is almost identical with the parameter a of the Poisson distribution. In Table 2 we wrote for every poem first the distribution whose P was greater.

Thus the only model expressing the word-length behaviour of Bachletová's poetry is the binomial distribution with its limiting case, the Poisson distribution ($n \to \infty$, $p \to 0$, $np \to a$). The result shows that the author has a certain "casting-mould" represented by a restricted <*I,S*> domain. It should be mentioned that for other texts, other variables and other languages, the <*I,S*> criterion yields very different results (c.f. Popescu et al. 2009) which may turn out to be characteristic of the author, style or language, etc.

3. Verse length

In some poetry the length of the verse measured in terms of word numbers is a constant. In such cases the bounding is too strong and verse length is not a variable. The same may hold for the number of feet (e.g. hexameter), number of syllables (e.g. thirteen), etc. However, in the poems by Bachletová which are free of any binding, verse length is a variable, most probably applied according to an internally pre-formed pattern and uttered spontaneously. Though the poems are short, the <*I,S*> characterization is always possible and if there are also longer verses, a probability distribution may be found.

Not all of the poems could be analyzed. Some of them were very short and the representation of individual frequency classes was far from being reliable. We selected poems having at least 15 verses and at least 4 frequency classes, and obtained the results presented in Table 3. The distribution is slightly more complex than the Poisson. Starting from Wimmer-Altmann's general theory (2005), the Poisson distribution follows from the simple difference equation

$$(4) \qquad P_x = \left(1 + a_0 + \frac{a_1}{x}\right) P_{x-1}$$

where a_0 = -1 and $a_1 = a$, yielding the formula presented above. For verse lengths in non-metric poetry this approach is not sufficient and a further modifying parameter must be added. We conjecture

.

$$(5) \qquad P_x = \left(1 + a_0 + \frac{a_1}{x + b - 1}\right) P_{x-1}$$

whose solution (setting again a_0 = -1, $a_1 = a$) yields

$$(6) \qquad P_x = \frac{a^x}{b^{(x)}\,{}_1F_1(1;b;a)}, \quad x = 0,1,2,\ldots,$$

where $b^{(x)} = b(b+1)\ldots(b+x-1)$ is the ascending factorial function, and ${}_1F_1(1;a;b)$ is the confluent hypergeometric function yielding the normalizing sum. Of course, since the distributions do not have $x = 0$, the expression (6) must be displaced one step to the right, i.e.

$$(7) \qquad P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1(1;b;a)}, \; x = 1,2,3,\ldots$$

an operation made by the software (Fitter) automatically. This is the one-displaced Hyperpoisson distribution. In one case, namely with the poem *Čas pre nádych vône* which does not have verses consisting of only one word, the distribution is displaced two step to the right.

In some cases the Poisson distribution which is a special case of Hyperpoisson (when $b = 1$) would be sufficient and in one case we were forced to use the limiting case of the Hyperpoisson, namely the geometric distribution, following for $a \to \infty$, $b \to \infty$, $a/b \to q$, where q is the parameter of the geometric distribution $P_x = pq^x$, $x = 0,1,2\ldots$

Table 3
Verse lengths in terms of word numbers

Poem	**Frequencies**	**Parameter**		$\mathbf{X^2}$	**DF**	**P**	**I**	**S**
		a	**b**					
Aby spriesvitela	4,15,4,3,1	0,6658;	0,1776	2,39	1	0,12	0,41	0,92
Bez rozlúčky	4,6,5,1	0,8294;	0,4813	0,73	1	0,39	0,36	0,15
Čakanie na boží jas	7,6,10,4,1,0,1	1,9925;	1,5222	2,39	2	0,30	0,71	1,31
Čas pre nádych vône	0,1,3,4,2,5,2	2,6402;	0,8863	1,85	3	0,60	0,58	-0,21

Hľadanie odpovedí	5,4,6,9	21,9759;27,4687	1,64	1	0,20	0,48	-0,46
Iba neha	13,15,13,9,3,1	1,9655; 1,4877	0,78	3	0,85	0,63	0,65
Ihly na nebi	2,12,3,1,3	0,7178; 0,1196	2,80	1	0,09	0,54	1,24
Malý ošiaľ	2,14,6,5	0,7569; 0,1081	1,22	1	0,27	0,31	0,38
Moje určenie	10,15,12,9,2,4	2,4553; 1,6677	1,93	3	0,59	0,73	0,94
Nepoznateľné	26,16,2,6,1	3,9217; 6,6769	6,96	2	0,03	0,64	1,39
Podobnosť bytia	4,6,10,7,1,1	1,8377; 0,8675	2,65	3	0,45	0,49	0,30
Rozdelená bytosť	1,9,8,5,2,1	1,2280; 0,1534	0,17	2	0,92	0,44	0,77
Som iná	1,7,10,2,1	0,9604; 0,1372	1,96	2	0,37	0,27	0,42
Stály smútok pre šesť písmen	4,13,13,15,2,1	1,6148; 0,4969	5,41	3	0,14	0,42	0,15
Tak málo úsmevu	1,3,10,4,2	1,6701; 0,5567	4,53	2	0,10	0,29	0,03
Vo večnosti slobodná	8,22,15,6,5,2,2	2,1271; 1,3296	4,20	3	0,24	0,75	1,50
Vyznania	7,13,4,0,2	0,5561; 0,2994	0,51	1	0,48	0,52	1,43
Z neba do neba	22,12,5,1	0,8735; 1,5399	0,20	1	0,65	0,39	0,85

The poem *Nepoznateľné* can be captured rather using the geometric distribution with parameter p = 0,5517 (X^2 = 1,33, DF = 1, P = 0,27) or using the Poisson distribution with parameter a = 0,7134, (X^2 = 0,31, DF = 1, P = 0,57). All the other data can be well fitted using the Hyperpoisson. Only in one case (*Hľadanie odpovedí*) the parameters seem to increase but there is no necessity to use a different distribution.

Looking at the <*I*,*S*> domain we see that verse-lengths are placed in the same domain as word-lengths. However, here the dispersion is still greater than with words, hence everything points to the existence of an ellipsis

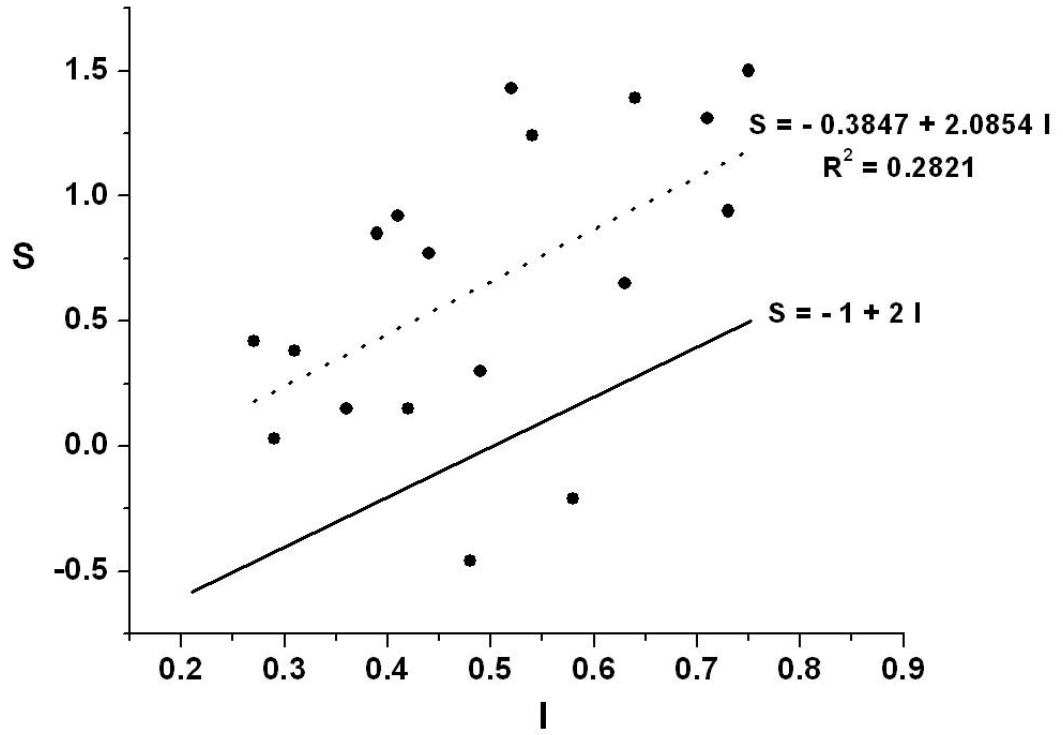

Figure 2. The <I,S> domain of verse length distributions

4. Verse length and Menzerath's law

According to Menzerath's law "the greater is a construct, the smaller are its components". The components are always immediate constituents. The law is stochastic and there is no transitivity, i.e. it does not hold necessarily that the greater the construct, the greater the components of the components. But just this is the contents of the so called Arens' law concerning sentence length and word length. This relationship evokes many problems and Grzybek, Stadlober and Kelih (Grzybek, Stadlober 2007; Grzybek, Stadlober, Kelih 2006, 2008) have shown that the result depends on text sort.

Since Bachletová poetry has its singular character (no rhyme, no fixed verse length, no meter) expressed especially by the shortness of verses, we can consider verse as a poetic construct whose immediate components are words. Bachletová's verse is a poetic substitute for the linguistic clause. If this conjecture is correct, then it must hold that the longer the verse, the shorter are its words on the average. The relationship follows from a simple logic: if the poet puts only one word in the verse, then it is most probably an autosemantic, and autosemantics are usually longer than synsemantics; but if he prolongs the verse, he has the chance to insert short synsemantics between autosemantics whereby the mean word length decreases. If our hypothesis is correct, we

have two problems: to test the hypothesis on individual poems and to obtain the parameters of this relationship characteristic for Bachletová.

This hypothesis is easily testable but the result strongly depends on the representativeness of verse numbers of a certain length. It is not testable using short poems. In its present form the hypothesis is an analogue to Menzerath's hypothesis but does not concern directly linguistic constructs but rather poetic, textual ones.

To test the above hypothesis we have chosen 14 longer poems. The results are presented in Table 4. Though not all length classes were representative, the general power trend representing Menzerath's law, $y = ax^{-b}$, could be shown. The parameters and the determination coefficients are shown in the last three columns of Table 4.

Table 4
Menzerath's law for verses

Poem	**Verse length (Number of words)**								**a**	**b**	**R^2**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**			
Aby spriesvitnela	3,5	2,63	1,92	1,83	1,4				3,5534	-0,5215	0,98
Iba neha	2,54	2,22	1,58	1,64	1,6	1,5			2,5586	-0,3147	0,91
Čakanie na boží jas	2,38	1,8	1,97	1,75	1,6	1,5			2,3303	-0,2281	0,86
Hladanie odpovedí	3	2,38	2,04	1,79					2,4823	-0,0174	0,80
Idem za tebou	---	2,5	2,5	1,88	1,9	1,89	1,71	1,5	3,3009	-0,3446	0,86
Moje určenie	3,3	2,17	1,93	1,68	1,7	1,5			3,2094	-0,4466	0,97
Rozťatá pritomnosť	2,67	1,83	1,71	1,25					2,6658	-0,4920	0,96
Tak málo úsmevu	5	1,83	2,1	2	1,9				4,6872	-0,7409	0,81
Môj ošiaľ	2,5	1,89	1,67	1,5					2,4883	-0,3703	~1.0
Podobnosť bytia	2,75	2,57	2,19	2	2,2	1,5			2,8539	-0,2487	0,78
Nepoznateľné	2,89	1,53	1,67	1,38	1,2				2,7772	-0.5479	0,90
Dielo Stvoriteľa	3,5	2,54	2,06	1,97	1,67	1,33			3,4948	-0,4542	0,99
Z neba do neba	2,73	2,13	1,78						2,7380	-0,3813	~1.0
Rozdelená bytosť	4,00	2,59	1,53	1,80	1,60	1,33			3,9520	-0,6328	0,95

The result has some textological consequences. We see that in text still other constructs than the usual linguistic ones abide by some regularities. In our case it was the verse but it can be asked whether there are still other entities, which underlie stochastic regularities, e.g. the strophe, the rhymed pair of verses, semantically or associatively constructed images of reality, etc. The setting up of such units and finding the relevant regularities holding for them are very demanding tasks whose solution must be delayed to future research.

References

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2006). *Quantitative Linguistik. Eine Annäherung*. Göttingen: Peust & Gutschmidt.

Grzybek, P. (ed.) (2006). *Contributions to the science of text and language. Word length studies and related issues.* Dordrecht: Springer.

Grzybek, P. (2006a). History and methodology of word length studies. In: Grzybek (2006): *15-90.*

Grzybek, P., Stadlober, E. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 205-217.* Berlin-New York: Mouton de Gruyter.

Grzybek, P., Stadlober, E., Kelih, E. (2006) The relationship of word length and sentence length. The inter-textual perspective. In: Decker, R., Lenz, H.-J. (eds.), *Advances in Data Analysis. Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8-10, 2006: 611-618*. Berlin, Heidelberg: Springer

Grzybek, P., Stadlober, E., Kelih, E. (2008). The relation between word Length and sentence length: an intra-systemic perspective in the core data structure. *Glottometrics 16, 111-121.*

Köhler, R. (2005). Language synergetics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin-New York: de Gruyter.

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Rietveld, T., Hout, R.v., Ernestus, M. (2004). Pitfalls in corpus research. *Computers and the Humanities 38(4), 343-362.*

Uhlířová, L. (1995). On the generality of statistical laws and individuality of texts. A case of syllables, word forms, their length and frequencies. *Journal of Quantitative Linguistics 2, 238-247.*

Wilson, A. (2006). Word-length distribution in present-day Lower Sorbian newspaper texts. In: Grzybek (2006): *319-327.*

Wimmer, G., Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 307-316*. Berlin-New York: de Gruyter.

Wimmer, G., Witkovský, V., Altmann, G. (1999). Modification of probability distributions applied to word length research. *Journal of Quantitative Linguistics 6, 257-268.*

Zipf, G.K. (1935/1968). *The psycho-biology of language: an introduction to dynamic philology.* Cambridge, Mass.: The M.I.T. Press.

Glottometrics 22, 2011, 57-61

Word length distribution in French

Karl-Heinz Best

Abstract. The present paper deals with word length in French letters and is part of the Göttingen *Projekt Quantitative Linguistik*. According to the examinations the *Hirata-Poisson distribution* yields good results. Hence the paper yields a further corroboration of the hypothesis that word length distributions abide by laws.

1. The state of the art

Word length is one of the most investigated phenomena in the Göttingen *Projekt Quantitative Linguistik* (Best 1998). The examinations are based on the theoretical works of Wimmer et al. (1994) and Wimmer & Altmann (1996) (cf. also Best 2006: 27ff.). The theory presented there, namely that the probability of appearance of words of length *x* is proportional to that of *x*-1, has been meanwhile corroborated on more than 4000 texts in about 50 languages. Our aim is to add some data concerning French.

2. Data and procedures

The presentation is based upon an examination of Knopp (1998) who evaluated letters of Friedrich II, Leibniz and Voltaire. Knopp measured word length in terms of syllable numbers in the word; the criterion of syllable number was the number of vowels in the word. The individual letters were examined as to their agreement with the 1-displaced Hirata-Poisson distribution (Wimmer, Altmann 1999: 256f.) which is identical with the Hermite distribution and arises both as a compounding and Feller-generalization as well as a convolution of the Poisson distribution:

$$P_x = \sum_{i=0}^{\left\lfloor \frac{x-1}{2} \right\rfloor} \binom{x-1-i}{i} \frac{e^{-a}\, a^{x-1-i}}{(x-1-i)!} b^i (1-b)^{x-1-2i}, \quad x = 1, 2, \ldots$$

where $\lfloor . \rfloor$ represents the integer part of the given number. The fitting of this distribution was successful in all cases. Thus the 1-displaced Hirata-Poisson has been repeatedly corroborated as an adequate model of French word length distributions (cf. Dieckmann & Judt 1996; Feldt, Janssen & Kuleisa 1997, Heinicke 2008).

In this paper the letters will not be analyzed separately but those of individual authors will be pooled and it will be tested whether these new samples follow the same regularity. At the same time, this is a test of the homogeneity of the style of individual authors.

3. Results

The fitting of the 1-displaced Hirata-Poisson distribution to the word lengths in Leibniz' letters yielded the results presented in Table 1. The letters were written to different addressees and originate from the years 1688-1696.

Table 1
Word lengths in 20 letters of Leibniz to different addressees

x	n_x	NP_x	x	n_x	NP_x
1	5695	5698.54	4	416	392.13
2	2764	2798.95	5	92	110.91
3	1309	1256.61	6	15	33.86
a = 0.5911		b = 0.1690			C = 0.0017

Here x = word length (in terms of syllable numbers); n_x = observed frequencies of words with x syllables; NP_x = frequencies computed by means of the Hirata-Poisson distribution; a, b = parameters; $C = X^2/N$, the discrepancy coefficient. It is used mostly with large samples and $C \leq 0{,}01$ is considered a good agreement with the model. This condition is fulfilled in all cases.

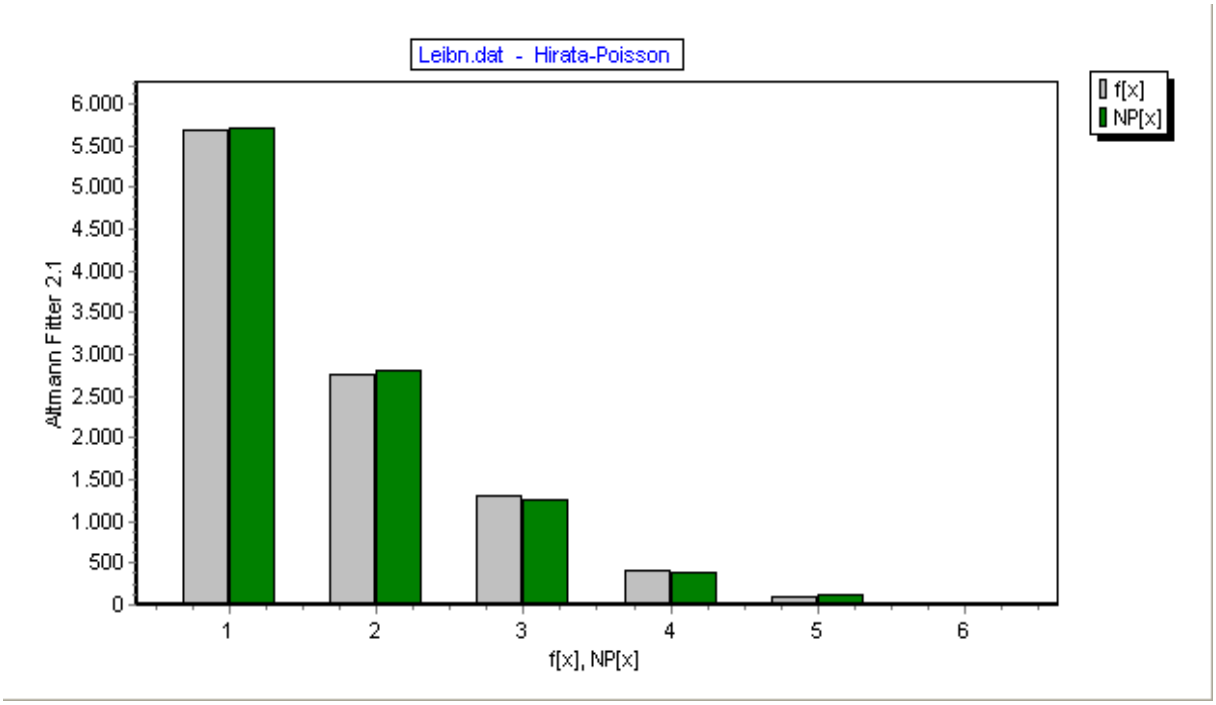

Figure 1. Word lengths in 20 letters of Leibniz to different addressees (the bright columns represent the observed values, the dark ones the computed ones)

The fitting of the 1-displaced Hirata-Poisson distribution to the letters of Friedrich II using the *Altmann-Fitter* (1997) yielded the results presented in Table 2 and Figure 2. The letters are addressed to Voltaire and were created in the years 1736-1742.

Table 2
Word length in 20 letters of Friedrich II to Voltaire

x	n_x	NP_x	x	n_x	NP_x
1	8057	8067.64	4	570	519.03
2	3390	3439.99	5	101	158.19
3	1819	1706.19	6	4	49.96
a = 0.5470		b = 0.2204			C = 0.0055

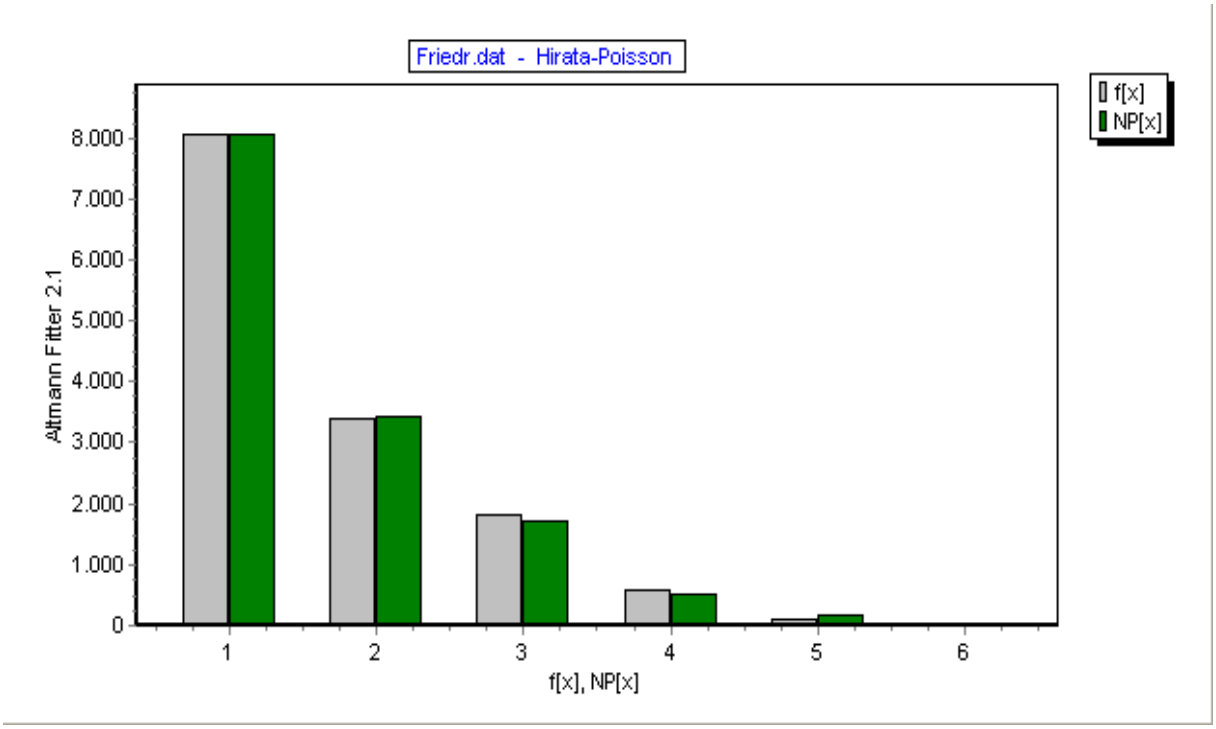

Figure 2. Word lengths in 20 letters of Friedrich II to Voltaire

For Voltaire's letters addressed to Friedrich II written in years 1736-1749 the results presented in Table 3 and Figure 3 were obtained.

Table 3
Word lengths in 20 letters of Voltaire to Friedrich II

x	n_x	NP_x	x	n_x	NP_x
1	8383	8559.85	4	500	533.47|
2	3833	3819.03	5	64	153.43|
3	2092	1763.65	6	4	46.57|
a = 0.5527		b = 0.1927			C = 0.0068

The vertical lines in the Tables mark the pooling of classes.

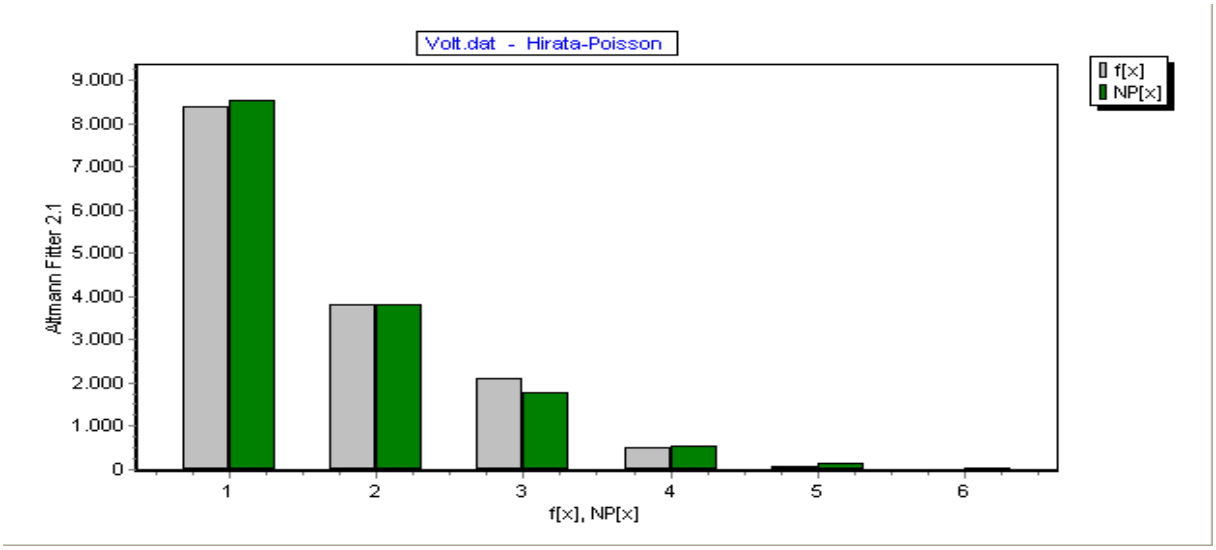

Figure 3. Word lengths in 20 letters of Voltaire to Friedrich II.

4. Excursus: Word length averages in the three text groups

The survey of average word lengths is presented in Table 4.

Table 4
Word length averages

Letters of	No of words	Average
Leibniz	10291	1.69
Friedrich II	13941	1.66
Voltaire	14876	1.66

As can be seen, the averages are almost identical with all authors.

5. Summary

The examination of Knopp (1998) has shown that to all of 60 individual letters the 1-displaced Hirata-Poisson distribution could be fitted.

In this paper, we have shown that even if one sets up samples of all letters of individual authors, the fitting is satisfactory.

This results obtained from French texts corroborate again the hypothesis of Wimmer & Altmann (1996) and Wimmer et al. (1994) that word lengths behave according to a background law. Based on Altmann's (1988) conjectures it can be supposed that the same theory holds true also for sentence lengths; the results obtained in Göttingen *Projekt Quantitative Linguistik* enable us to conclude that also the distributions of morphs (Best 2005b), rhythmic units (Best 2005a) and syllables (Cassier 2001) abide by laws.

Texts

The Complete Works of Voltaire. Oxford: The Voltaire Foundation [at the] Taylor Institution 1986 (contains letter written by Voltaire and addresses to him).

Leibniz, Gottfried Wilhelm (1970). *Sämtliche Schriften und Briefe*. Band 5-13. Berlin: Akademie-Verlag.

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Best, K.-H. (1998). Results and perspectives of the Göttingen Project on Quantitative Linguistics. *Journal of Quantitative Linguistics 5, 155-162.*

Best, K.-H. (2005a). Längen rhythmischer Einheiten. In: Köhler, R., Altmann, G., Piotowski, R.G. (eds..), *Quantitative Linguistik - Quantitative Linguistics. An International Handbook: 208-214*. Berlin-NewYork: de Gruyter.

Best, K.-H. (2005b). Morphlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik - Quantitative Linguistics. An International Handbook: 255-260*. Berlin-NewYork: de Gruyter.

Best, K.-H. (2006). *Quantitative Linguistik: Eine Annäherung.* 3rd revised and extended edition. Göttingen: Peust & Gutschmidt.

Cassier, F.-U. (2001). Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 33-42.* Göttingen: Peust & Gutschmidt.

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In: Schmidt, P. (ed.), *Glottometrika 15, 158-165.* Trier: Wissenschaftlicher Verlag Trier.

Feldt, S., Janssen, M., & Kuleisa, S. (1997). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In: Best, K.-H. (ed.), *Glottometrika 16, 145-151*. Trier: Wissenschaftlicher Verlag Trier.

Heinicke, N. (2008). Wortlängenverteilungen in französischen Briefen eines Autors. *Glottoetrics 16, 38-45.*

Knopp, A. (1998). *Wortlängen in französischen Briefen deutschsprachiger Verfasser*. Göttingen, Staatsexamensarbeit.

Wimmer, G., Altmann, G. (1996). The theory of word length distribution: some results and generalizations. In: Schmidt, P. (ed.), *Glottometrika 15, 112-133*. Trier: Wissenschaftlicher Verlag Trier.

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

Wimmer, G, Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Software

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Informations about the Göttingen *Projekt Quantitative Linguistik*: Homepage:
http://wwwuser.gwdg.de/~kbest/

Glottometrics 22, 2011, 62-72

Vocabulary richness in Slovak poetry

Ioan-Iovitz Popescu
Radek Čech[1]
Gabriel Altmann

Abstract. This article examines different indicators of text properties – such as entropy, repeat rate, and arc length – and their distribution. All of these can be described as indicators of the vocabulary richness of the texts, as there is a very strict linear relationship between them.

Keywords: *vocabulary richness, rank-frequency distribution, entropy, repeat rate, Gini's coefficient*

1. Introduction

The study of vocabulary richness has had a long tradition in linguistic studies focused on the frequency characteristics of texts. The majority of proposed approaches have struggled with the impact of text length on vocabulary size (cf. Baayen 1989; Bernett 1988; Covington, McFall 2010; Ejiri, Smith 1993; Guiraud 1954, 1959; Herdan 1960, 1966; Hess, Sefton, Landry 1986, 1989; Honore 1979; Martynenko 2010; Menard 1983; Müller D. 2002; Panas 2001; Popescu et al. 2009; Popescu, Čech and Altmann 2011a, 2011b; Ratkowsky, Hantrais 1975; Tešitelová 1972; Tuldava, 1995; Tuzzi, Popescu and Altmann 2010; Tweedie, Baayen 1998; Weitzman 1971; Yule 1944 – to mention only some of the relevant studies). It is obvious that in order to achieve an appropriate measurement of vocabulary richness it is necessary to eliminate the detrimental factor of text length by means of some transformation. Further, as has been shown by Thoiron (1986) and Popescu, Čech, Altmann (2011b), entropy and repeat rate can also be used to measure vocabulary richness.

In this paper we examine some indicators of vocabulary richness proposed earlier by Popescu et al. (2009) and Popescu, Čech and Altmann (2011a, 2011b), applying them to 54 poems by the Slovak writer Eva Bachletová. Moreover, a new indicator is introduced. In this way one can obtain an overall picture of one of the many aspects of poetic creativity.

Clearly if the poems are short, few words are repeated and the text seemingly displays a high degree of vocabulary richness. The situation changes if the text becomes longer. The frequency of repeated words increases more rapidly than the number of unique words (hapax legomena). Nevertheless, hapaxes would continue to appear despite the length of texts, but if the texts become very long, the rate of occurrence of new words would drop. Text length thus affects the data. The meaning of 'short' and 'long' texts has never been precisely defined. In statistics, 'long' means infinite, but with some classical tests it begins with $N = 120$. With some other tests, e.g. the chi-square, the more cases there are, the worse the result (cf. Rietveld, Hout, Ernestus 2004); this holds only for data sets not too large and not too small, but this is difficult to determine.

[1] Address correspondence to: Radek Čech, Department of Czech Language, University of Ostrava, Reální 5, Ostrava, 701 03, Czech Republic, e-mail: radek.cech@osu.cz,

Thus, if one establishes an indicator of vocabulary richness, one has only a unique criterion for measuring its goodness, viz. its strong correlation with some other indicators interpreted as expressions of this property.

2. Gini's coefficient

If we compute the rank-frequency distribution of word forms of a text and reverse the ranking, i.e. if we begin to rank the frequencies 'from below', then the cumulative relative frequencies form a curve called the Lorenz curve, which for word frequencies is always placed below the $x = y$ line (the bisector of the first quadrant), whereby also the ranks must be relativized, i.e. $x, y \in \langle 0, 1 \rangle$. The area between the bisector and the Lorenz curve is usually called Gini's coefficient, as can be seen in Figure 1.

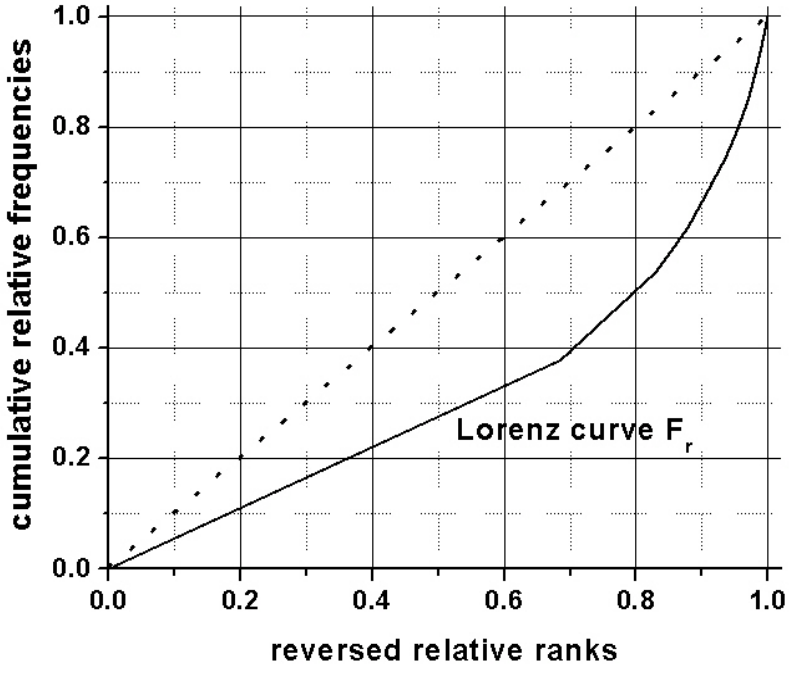

Figure 1. The Lorenz curve (from Popescu, I.-I. et al. 2009: 57)

For its computation without the reversion of ranks and cumulation, one uses the equivalent expression

$$(1) \qquad G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{r=1}^{V} rf(r)\right) = \frac{1}{V}\left(V + 1 - 2m_1'\right),$$

in which the last expression ($2m_1'$) is twice the mean of the rank-frequency distribution, V is the highest rank (number of word types), and N is the number of tokens, i.e. text length. Since

the greater the area between the bisector and the Lorenz curve, the smaller the vocabulary richness, as shown in Popescu et al. (2009: 57), the authors propose a complementary indicator

(2) $R_4 = 1 - G$.

which shows richness directly. In order to illustrate the procedure, we compute Gini's coefficient for the short poem *Bez rozlúčky* as presented in Table 1.

Table 1
Rank-frequency distribution of word forms
in E. Bachletová's poem *Bez rozlúčky*

Rank r	**Frequency f(r)**		**Rank r**	**Frequency f(r)**
1	2		17	1
2	2		18	1
3	2		19	1
4	1		20	1
5	1		21	1
6	1		22	1
7	1		23	1
8	1		24	1
9	1		25	1
10	1		26	1
11	1		27	1
12	1		28	1
13	1		29	1
14	1		30	1
15	1		31	1
16	1		32	1

Here $V = 32$, $N = 35$. The mean can easily be computed as

$$m_1' = [1(2) + 2(2) + 3(2) + 4(1) + \ldots + 32(1)]/35 = 15.2571.$$

Inserting these numbers into formula (1) we obtain

$$G = \frac{1}{32}\left(32 + 1 - (2)15.2571\right) = 0.0777$$

Hence $R_4 = 1 - G = 1 - 0.0777 = 0.9223$. All values of G and R_4 concerning individual poems by E. Bachletová are presented in Table 2. They are ordered according to increasing N. As can

easily be seen in Table 2, here *G* does not depend on *N*, an important property of text indicators. Nevertheless, it is possible that very large *N* can destroy this advantage.

Table 2
Gini's coefficient and the richness indicator R_4

Poem	**N**	**G**	**R_4**	**Var(G)**
Miesto pre Nádej	29	0.0333	0.9667	0.0122
Ťažko pokoriteľní	30	0.1205	0.8795	0.0128
Tiché verše	31	0.0601	0.9399	0.0117
Ulomené zo slov	31	0.1600	0.8400	0.0122
Dovoľ mi slúžiť	34	0.0285	0.9715	0.0103
Len áno	34	0.1525	0.8475	0.0105
Bez rozlúčky	35	0.0777	0.9223	0.0105
Pravidlá odpúšťania	35	0.1069	0.8931	0.0110
Tá Láska	35	0.1190	0.8810	0.0106
Dnešný luxus	36	0.1925	0.8075	0.0109
Neopusť ma...	36	0.3363	0.6637	0.1120
Zbytočné srdce	36	0.2202	0.7798	0.0111
Vďaka Pane!	37	0.0510	0.9490	0.0097
Nado mnou Ty sám...	38	0.1106	0.8894	0.0101
Vďaka za deň	39	0.0705	0.9295	0.0094
Istota	41	0.1729	0.8271	0.0096
Ešte raz	42	0.1890	0.8110	0.0094
Iba život	44	0.0444	0.9556	0.0082
Kým ich máme	44	0.1072	0.8928	0.0088
Večerná ruža	46	0.0425	0.9575	0.0078
Čakáme šťastie...	48	0.0979	0.9021	0.0079
Spájania	48	0.0979	0.9021	0.0079
Do večnosti beží čas	51	0.1917	0.8083	0.0078
Malé modlitby	51	0.1461	0.8539	0.0074
Precitnutie	51	0.0908	0.9092	0.0074
Vrátili sa	51	0.0908	0.9092	0.0074
Keď dohorí deň	52	0.1592	0.8408	0.0076
Zasľúbenie jasu	52	0.1726	0.8274	0.0073
Ihly na nebi	54	0.2270	0.7730	0.0070
Vyznania	55	0.0994	0.9006	0.0069
Naše mamy	56	0.1173	0.8827	0.0069
Som iná	58	0.2285	0.7715	0.0067
To všetko je dar	58	0.2967	0.7033	0.0067

Aby spriesvitnela	63	0.1450	0.8550	0.0062
Tak málo úsmevu	63	0.1484	0.8516	0.0064
Hľadanie odpovedí	67	0.1176	0.8824	0.0056
Naše svetlo	67	0.2604	0.7396	0.0059
Z neba do neba	67	0.1661	0.8339	0.0060
Malý ošiaľ	68	0.2699	0.7301	0.0055
Večerné ticho	68	0.1992	0.8008	0.0059
Idem za Tebou	72	0.0893	0.9107	0.0052
Čakanie na Boží jas	77	0.2157	0.7843	0.0053
Rozťatá prítomnosť	78	0.1944	0.8056	0.0049
Rozdelená bytosť	79	0.1022	0.8978	0.0048
Čas pre nádych vône	81	0.0816	0.9184	0.0046
Prvotný sen	81	0.0961	0.9039	0.0047
Podobnosť bytia	85	0.1459	0.8541	0.0047
Náš chrám	86	0.1554	0.8446	0.0047
Nepoznateľné	93	0.2300	0.7700	0.0043
Dielo Stvoriteľa	136	0.1566	0.8434	0.0029
Iba neha	139	0.2757	0.7243	0.0028
Moje určenie	146	0.1896	0.8104	0.0027
Stály smútok pre šesť písmen	146	0.3118	0.6882	0.0027
Vo večnosti slobodná	170	0.2330	0.7670	0.0024

G or R_4 have the advantage of allowing for an easy comparison of texts. Looking at G or R_4 in formula (1), where V is a constant, we can state that the asymptotic variance is given by

$$Var(G) = \frac{4}{V^2} Var(m_1') = \frac{4m_2}{V^2 N} \tag{3}$$

where m_2 is the variance of the distribution. The variance of R_4 is identical because 1 is a constant. All variances are presented in the last column of Table 2.

In order to compare two texts, one can perform an asymptotic normal test using the criterion

$$u = \frac{|G_1 - G_2|}{\sqrt{Var(G_1) + Var(G_2)}}, \tag{4}$$

where the subscript numbers 1 and 2 refer to two different texts. For example, comparing the first and the last text in Table 2 we obtain

$$u = \frac{|0.0333 - 0.2330|}{\sqrt{0.0122 + 0.0024}} = 1.65$$

which, in a two-sided test, is not significant. Even the greatest difference of *G* existing between the poems *Miesto pre Nádej* and *Neopusť ma* is not significant. Hence we can state that the author has a special technique of using her vocabulary in her poems.

3. Arc length, Repeat rate and Entropy

As has been said above, a satisfactory indicator of vocabulary richness must correlate with other indicators expressing the same quality. In a previous article (Popescu, Čech, Altmann 2011b) we presented the indicator R_1, the relative entropy H_{rel} and the relative repeat rate $RR_{McIntosh}$. Here we add the indicator *R*, whose computation is somewhat more complex mathematically but is nevertheless straightforward using a computer. This indicator expresses richness from a different point of view: it is based on the two parts of the arc joining the frequency at the first rank $f(1)$ and at the last rank $f(V)$. The arc *L* is defined as the sum of Euclidean distances between neighbouring frequencies, i.e.

$$L = \sum_{r=1}^{V-1} \{[f(r) - f(r+1)]^2 + 1\}^{1/2}. \tag{5}$$

For example, for the distribution in Table 1 we obtain

$$L = [(2-2)^2 + 1]^{1/2} + [(2-2)^2 + 1]^{1/2} + [(2-1)^2 + 1]^{1/2} + [(1-1)^2 + 1]^{1/2} + \ldots + [(1-1)^2 + 1]^{1/2} = 31.4142.$$

The *h*-point is defined as

$$h = \begin{cases} r, & \text{if there is an } r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \text{if there is no } r = f(r) \end{cases} \tag{6}$$

i.e. that point at which $r = f(r)$, or, if there is no such point, it is computed by means of the second part of formula (6). In the first case, *h* is an integer; in the second case it is a positive real number.[2] This point has been used directly for computing the richness indicator R_1 (cf.

[2] In scientometrics it is called Hirsch's index or h-index (Hirsch 2005); it has been introduced to linguistics by Popescu (2007).

Popescu et al. 2009: 33)[3]; here we use it to compute that part of the arc length which is above the *h*-point in order to set up the indicator

$$R = 1 - \frac{L_h}{L} \tag{7}$$

The computation of L_h is straightforward if *h* is an integer. However, if it has a positive real value, we must add to the arc up to [*h*] that part of the arc which lies between the integer part of *h* (= [*h*]) and *h* itself, i.e. we compute

$$L_h = \sum_{r=1}^{[h]-1} \{[f(r) - f(r+1)]^2 + 1\}^{1/2} + \{(h - f([h]))^2 + (h - [h])^2\}^{1/2}. \tag{8}$$

In order to illustrate this computation, imagine a distribution of the following form

r	*f*(*r*)
1	5
2	3
3	1
…….	

Evidently, the *h*-point is between *r* = 2 and *r* = 3, and we compute it using the second part of formula (5) as

h = [3(3)-2(1)]/[3 – 2 +3 – 1] = 7/3 = 2.3333.

Hence L_h consists of $[(5 - 3)^2 + 1]^{1/2} + [(2{,}3333 - 3)^2 + (2{,}3333 - 2)^2]^{1/2} = 2.9814$.

In Table 3 we show all indicators together and compare *R* with the others, namely (a) R_1 containing *F*(*h*), *h* and *N* (see footnote 2); (b) the repeat rate relativized according to McIntosh (*RRmc*)

$$RR_{mc} = \frac{1 - \sqrt{RR}}{1 - 1/\sqrt{V}},$$

where V is the number of types (vocabulary); (c) the relative entropy H_{rel}

$$H_{rel} = \frac{H}{H_0};$$

and (d) R_4 = 1-*G* using Gini's coefficient.

---

[3] $R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right)$, where *F*(*h*) is the sum of relative frequencies from *r* = 1 up to *r* = [*h*]. Since *h* may be a positive real number, we subtract from *F*(*h*) the relativized half of the square built by *h*, i.e. we add this part to 1-*F*(*h*)..

Table 3
Survey of some richness indicators applied to poems by E. Bachletová

Poem	**R**	**R_1**	**RR_{mc}**	**H_{rel}**	**R_4**
Aby spriesvitnela	0.9547	0.9127	0.9818	0.9783	0.8550
Bez rozlúčky	0.9682	0.9429	0.9925	0.9916	0.9223
Čakáme šťastie...	0.9767	0.9401	0.9864	0.9851	0.9021
Čakanie na Boží jas	0.8972	0.8506	0.9510	0.9521	0.7843
Čas pre nádych vône	0.9865	0.9645	0.9925	0.9902	0.9184
Dielo Stvoriteľa	0.9524	0.9228	0.9797	0.9751	0.8434
Dnešný luxus	0.9499	0.8924	0.9677	0.9670	0.8075
Do večnosti beží čas	0.9400	0.8725	0.9706	0.9673	0.8083
Dovoľ mi slúžiť	1	0.9743	0.9971	0.9968	0.9715
Ešte raz	0.9274	0.8690	0.9696	0.9674	0.8110
Hľadanie odpovedí	0.9755	0.9552	0.9909	0.9878	0.8824
Iba neha	0.9194	0.8901	0.9603	0.9523	0.7243
Iba život	1	0.9520	0.9924	0.9924	0.9556
Idem za Tebou	0.9782	0.9583	0.9929	0.9905	0.9107
Ihly na nebi	0.9385	0.8981	0.9724	0.9661	0.7730
Istota	0.9575	0.9055	0.9727	0.9714	0.8271
Keď dohorí deň	0.9493	0.9062	0.9720	0.9713	0.8408
Kým ich máme	0.9436	0.9091	0.9810	0.9813	0.8928
Len áno	0.9621	0.9412	0.9860	0.9834	0.8475
Malé modlitby	0.9662	0.9412	0.9871	0.9838	0.8539
Malý ošiaľ	0.8932	0.8750	0.9607	0.9547	0.7301
Miesto pre Nádej	1	0.9698	0.9963	0.9962	0.9667
Moje určenie	0.9301	0.9075	0.9707	0.9674	0.8104
Nado mnou Ty sám...	0.9355	0.8947	0.9780	0.9799	0.8894
Náš chrám	0.9209	0.8968	0.9607	0.9637	0.8446
Naše mamy	0.9713	0.9308	0.9827	0.9810	0.8827
Naše svetlo	0.9284	0.8507	0.9589	0.9504	0.7396
Neopusť ma...	0.8665	0.8646	0.9384	0.9455	0.6637
Nepoznateľné	0.9253	0.8763	0.9636	0.9581	0.7700
Podobnosť bytia	0.9087	0.8941	0.9613	0.9654	0.8541
Pravidlá odpúšťania	0.968	0.9063	0.9802	0.9805	0.8931
Precitnutie	0.9691	0.9412	0.9904	0.9887	0.9092
Prvotný sen	0.9578	0.9275	0.9800	0.9798	0.9039
Rozdelená bytosť	0.9797	0.9620	0.9927	0.9897	0.8978
Rozťatá prítomnosť	0.9599	0.9295	0.9817	0.9750	0.8056

Som iná	0.9310	0.8716	0.9534	0.9536	0.7715
Spájania	0.9767	0.9401	0.9864	0.9851	0.9021
Stály smútok pre šesť písmen	0.9130	0.8493	0.9536	0.9407	0.6882
Tá Láska	0.9660	0.9429	0.9889	0.9871	0.8810
Tak málo úsmevu	0.8960	0.873	0.9537	0.9614	0.8516
Ťažko pokoriteľní	0.9452	0.9000	0.9817	0.9818	0.8795
Tiché verše	0.9648	0.9355	0.9937	0.9933	0.9399
To všetko je dar	0.9232	0.8170	0.9433	0.9350	0.7033
Ulomené zo slov	0.943	0.9032	0.9795	0.9782	0.8400
Vďaka Pane!	0.9709	0.9459	0.9952	0.9945	0.9490
Vďaka za deň	0.9718	0.9487	0.9936	0.9926	0.9295
Večerná ruža	1	0.9650	0.9929	0.9928	0.9575
Večerné ticho	0.9243	0.8897	0.9679	0.9638	0.8008
Vo večnosti slobodná	0.9544	0.8941	0.9716	0.9608	0.7670
Vrátili sa	0.9691	0.9412	0.9904	0.9887	0.9092
Vyznania	0.9710	0.9455	0.9905	0.9884	0.9006
Z neba do neba	0.9270	0.8881	0.9709	0.9692	0.8339
Zasľúbenie jasu	0.9463	0.9231	0.9812	0.9778	0.8274
Zbytočné srdce	0.8604	0.8333	0.9424	0.9500	0.7798

4. Relations

As can be seen in Table 3, whatever indicator we use, Bachletová's vocabulary richness is very high. The relationships are as follows:

$R = 0.2572 + 0.7580R_1$ with $R^2 = 0.78$
$R = - 0.7286 + 1.7209H_{rel}$ with $R^2 = 0.74$
$R = -0.8806 + 1.8732RR_{Mc}$ with $R^2 = 0.84$
$R = 0.6579 + 0.3416R_4$ with $R^2 = 0.58$.

All relations can be considered linear. In all cases we obtain highly significant values in *t*- and *F*-tests, even if the determination coefficient is not very high. We may conclude that *R* is an 'honest' indicator of vocabulary richness. Needless to say, further examinations using different texts in different languages will either corroborate or contradict this result, but in any case the individual parameters in the above equations will change if one adds more texts.

5. Conclusion

This article has presented a new indicator of vocabulary richness. The significant correlations with other indicators (see Table 3) allow us to assume that this indicator genuinely expresses the observed property of text. As for the measurement of vocabulary richness in general, we

are convinced that only a complex measurement based on different indicators can bring satisfactory results because the text is obviously a 'product' of a complex process controlled by different mechanisms. Moreover, all proposed indicators (each in its own way) eliminate the influence of the length of the text, which is the most problematic aspect of the measurement of vocabulary richness.

We assume that the method can not only be used for the measurement of vocabulary richness itself, but can also be used as an additional indicator in stylometrics.

Acknowledgment

Radek Čech was supported by the Czech Science Foundation, grant no. P406/11/0268.

References

Baayen, R.H. (1989). *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation*. Diss. Amsterdam: Free University.

Bernet, Ch. (1988). Faits lexicaux. Richesse du vocabulaire. In P. Thoiron, et al. (eds.), *Etudes sur la richesse et la structure lexicale: 1-11*. Paris: Champion.

Covington, M.A., & McFall, J.D. (2010). Cutting the Gordian Knot: The Moving Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics 17(2), 94-100.*

Ejiri, K., & Smith, A.E. (1993). Proposal of a new 'constraint measure' for text. In R. Köhler, B.B. Rieger, (Eds.), *Contributions to quantitative linguistics*: *195-211*. Dordrecht: Kluwer.

Guiraud, P. (1954). *Les catactères stitistiques du vocabulaire*. Paris: Presses Universitaires de France.

Guiraud, P. (1959). *Problèmes et methods de la statistique linguistique*. Dordrecht: Reidel.

Herdan, G. (1960). *Type-token mathematics*. The Hague: Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. New York: Springer.

Hess, C.E., Sefton, K.M., Landry, R.G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research 29, 129-134.*

Hess, C.E., Sefton, K.M., Landry, R.G. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research 32, 536-540.*

Honore, T. (1979). Some simple measures of richness of vocabulary. *ALLC Bulletin 7, 172-177.*

Martynenko, G. (2010). Measuring lexical richness and its harmony. In: P. Grzybek, E. Kelih, J. Mačutek, J. (Eds.), *Text and language*: *125-132.* Wien: Praesens.

Menard, N. (1983). *Mesure de la richesse lexicale*. Paris: Slatkine.

Müller, D. (2002). Computing the type token relation from the a priori distribution of types. *Journal of Quantitative Linguistics 9, 193-214.*

Panas, E. (2001). The generalized Torquist: Specification and estimation of a new vocabulary text-size function. *Journal of Quantitative Linguistics 8, 233-252.*

Popescu, I.-I. (2007)**.** Text ranking by the weight of highly frequent words. In: P. Grzybek, R. Köhler (Eds), *Exact methods in the study of language and text: 557-567.* Berlin – New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Čech, R., Altmann, G. (2011a). *The lambda-structure of texts.* Lüdenscheid: RAM.

Popescu, I.-I., Čech, R., Altmann, G. (2011b). Some characterizations of Slovak poetry. (Submitted)

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: **RAM.**

Ratkowsky, D.A., Hantrais, L. (1975). Tables for comparing the richness and structure of vocabulary in texts of different length. *Computers and Humanities 9, 69-75*.

Rietveld, T., Hout van, R., Ernestus, M. (2004). Pitfalls in Corpus Research. *Computers and the Humanities 38, 343–362.*

Tešitelová, M. (1972). On the so-called vocabulary richness. *Prague Studies in Mathematical Linguistics 3, 103-120*.

Thoiron, P. (1986). Diversity index and entropy as measures of vocabulary richness. *Computers and the Humanities 20, 197-202.*

Tuldava, J. (1995). On the relation between text length and vocabulary size. In: J. Tuldava (Ed.) *Methods in quantitative linguistics*: *131-150*. Trier: WVT.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts.* Lüdenscheid: RAM.

Tweedie, F.J., Baayen, R.H. (1998). How variable may a constant be? Measure of lexical richness in perspective. *Computers and the Humanities 32, 323- 352*.

Weizman, M. (1971). How useful is the logarithmic type-token ratio? *Journal of Linguistics 7, 237-243.*

Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Glottometrics 22, 2011, 73-85

Polysemy and word length in Chinese

Lu Wang, Trier

Abstract. This paper investigates the distribution of polysemy and the relationship between polysemy and word length in Chinese. The results show that polysemy in Chinese abides by the Zipf-Mandelbrot law; word length is, in accordance with other studies, a function of polysemy and vice versa.

Keywords: Polysemy, word length, Chinese

1. Introduction

In the past, polysemy in Chinese was mostly studied from a qualitative point of view whereas its quantitative features were seldom adopted into study so far. This paper aims at scrutinizing two of those features: the distribution of the polysemy of word types and the polysemy-length relation.

2. Hypotheses

It is well known that the number of meanings of words is lawfully distributed in the languages studied so far. Here, the vocabulary of a Chinese text corpus is used for further corroboration. We shall investigate two hypotheses.

Hypothesis 1: *The polysemy of Chinese lexemes is in accordance with the Zipf-Mandelbrot law.*

Since Zipf (1949) introduced the relationship between polysemy and word length into quantitative linguistics, many languages have been studied in this respect: German, Hungarian and Slovak by Altmann, Beőthy and Best (1982); German, Swedish and Indonesian by Fickermann, Markner-Jäger, Rothe (1984); Maori by Köhler (1999). All those research results agree with Altmann's (1989) assumption concerning compounds: *"The longer a compound the fewer meanings it has (on the average)"*. To test whether Zipf's law can be applied to Chinese, we set up the hypothesis 2:

Hypothesis 2: *In Chinese, word length* (*L*) *is a function of polysemy* (*P*) *following a power law and vice versa:*

$$L = aP^b + 1 \qquad (1)$$

$$P = cL^d + 1 \quad (2)$$

The constant 1 is added to the power function because both length and polysemy cannot be smaller than 1.

3. Data and method

In this study, we use the People's Daily Corpus January 1998, which is a Chinese one-million words news corpus with word segmentation and part-of-speech tagging. Firstly, we ignore all the words containing numbers or characters from alphabets (i.e. only words consisting of Chinese characters are taken into account); therefore, the number of word types shrinks from 82231 to 58742. Then, we look them up in the *Modern Chinese Dictionary (5th Edition)* to determine the number of meanings, which results in 22636 tagged words (omitting words which are not found in the dictionary). Finally, word length is measured in terms of the number of Chinese characters, which corresponds to the number of syllables.

The statistical result is shown in Table 1. The longest words contain 7 characters: 远水解不了近渴, 一失足成千古恨, 新民主主义革命, 马克思列宁主义, 车到山前必有路 and 不管三七二十一. All of them have polysemy one. On the other hand, the word with maximum polysemy is a one-character word:下. Its 28 meanings in the dictionary are:

下 [1]xià

(1) [名]方位词。位置在低处的：～游|～部|山～|往～看。

(2) 等次或品级低的：～等|～级|～策|～品。

(3) [名]方位词。次序或时间在后的：～次|～半年|～不为例。

(4) 向下面：～达|～行。

(5) [名]方位词。表示属于一定范围、情况、条件等：名～|部～|在党的领导～|在这种情况～。

(6) 表示当某个时间或时节：时～|节～|年～。

(7) 用在数目字后面，表示方面或方位：两～都同意|往四～一看。

(8) (Xià)[名]姓。

下 [2]xià

(1) [动]由高处到低处：～山|～楼|顺流而～。

(2) [动](雨、雪等)降落：～雨|～雪|～霜。

(3) [动]发布；投递：～命令|～通知|～战书。

(4) [动]去；到(处所)：～乡|～车间|～馆子。

(5) [动]退场：八一队的五号～，三号上|这一场戏你应该从右边的旁门～。

(6) [动]放入：～种|～面条|～本钱|～网捞鱼。

(7) [动]进行(棋类游艺或比赛)：～围棋|咱们～两盘象棋吧!

(8) [动]卸除；取下：～装|把敌人的枪～了|把窗户～下来。

(9) [动]做出(言论、判断等)：～结论|～批语|～定义。

(10) [动]使用；开始使用：～力气|～工夫|～刀|～笔|对症～药。

(11) [动](动物)生产：母猪～小猪|鸡～蛋。

(12) 攻陷：连～数城。

(13) 退让：相持不～。

(14) [动]到规定时间结束日常工作或学习等：～班|～课。

(15) [动]低于；少于：参加大会的不～三千人。

下 [3]xià

(～儿)[量]

(1) a)用于动作的次数：钟打了三～|摇了几～旗子。b)〈方〉用于器物的容量：瓶子里装着半～墨水|这么大的碗，他足足吃了三～。

(2) 用在“两、几”后面，表示本领、技能：他真有两～|就这么几～，你还要逞能?□也说下子。

下□·xià

[动]趋向动词。用在动词后。

(1) 表示由高处到低处：坐～|躺～|传～一道命令。

(2) 表示有空间，能容纳：坐得～|这个剧场能容～上千人|这间屋子太小，睡不～六个人。

(3) 表示动作的完成或结果：打～基础|定～计策|准备～材料。

Table 1
Polysemy, word length (in characters/syllables) and word types in the corpus.

Polysemy	**Word length**	**Word types**	**Polysemy**	**Word length**	**Word types**
1	1	47	6	2	13
1	2	12509	7	1	120
1	3	1223	7	2	3
1	4	1559	8	1	84
1	5	23	8	2	1
1	6	9	9	1	70
1	7	6	10	1	41
2	1	549	11	1	34
2	2	3730	12	1	21
2	3	182	13	1	9
2	4	68	14	1	8
2	5	1	15	1	7
3	1	506	16	1	5
3	2	707	17	1	6
3	3	16	18	1	2
3	4	1	19	1	3
4	1	386	21	1	3
4	2	128	24	1	1
4	3	2	25	1	1
5	1	313	26	1	1
6	1	196	28	1	1

4. Results

4.1. The Polysemy distribution

In Table 2 and Figure 1 we can see that the number of word types decreases rapidly with increasing polysemy. This means that the vocabulary of the corpus is mainly composed of words with very few meanings. Correspondingly, those with more meanings make up only a small proportion.

Table 2
Polysemy and number of word types.

Polysemy	Word types	Polysemy	Word types
1	15376	13	9
2	4530	14	8
3	1230	15	7
4	516	16	5
5	354	17	6
6	209	18	2
7	123	19	3
8	85	21	3
9	70	24	1
10	41	25	1
11	34	26	1
12	21	28	1

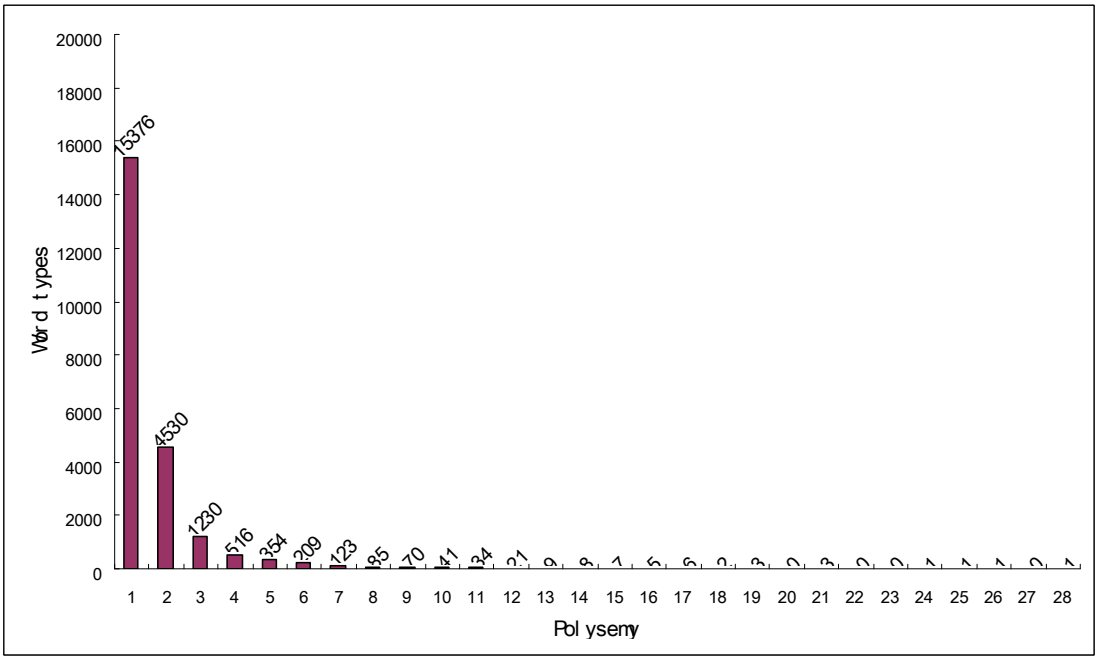

Figure 1. Empirical polysemy distribution in the corpus vocabulary.

As large amounts of data often prevent the application of the Chi-square test (its value increases with sample size) we use Zipf's and Zipf-Mandelbrot's law rather in form of a function, which does not depend on degrees of freedom. Among alternative models those with fewer parameters should be preferred (Occam's razor); that is why the simpler Zipf form is tested, too. However, as we will see, the results of fitting the functions to the data are not equivalent. The following three models are tested on the data from the corpus:

(3) $W = d + a/(b + P)^c$,

the Zipf-Mandelbrot model with an additional constant, where W is the number of word types and P the extent of polysemy; the usual power law (Zipf)

(4) $W = aP^{-b}$

and the stratificational view proposed by Popescu et al. (2010)

(5) $W = 1 + ae^{-P/b}$

The results are presented in Table 3.

Table 3
Fitting functions (3) to (5) to the distribution of polysemy in Chinese

Polysemy	**Number of lexemes**	**Zipf's law**	**Zipf-Mandelbrot's law**	**Popescu's law**
1	15376	15468.077	15389.331	15371.627
2	4530	3771.386	4387.361	4531.692
3	1230	1651.797	1481.498	1336.480
4	516	919.529	573.899	394.650
5	354	583.774	251.140	117.033
6	209	402.737	123.841	35.202
7	123	294.245	69.203	11.082
8	85	224.197	44.042	3.972
9	70	176.391	31.745	1.876
10	41	142.334	25.421	1.258
11	34	117.227	22.021	1.076
12	21	98.194	20.121	1.022
13	9	83.427	19.023	1.007
14	8	71.742	18.369	1.002
15	7	62.340	17.968	1.001
16	5	54.663	17.717	1.000
17	6	48.315	17.556	1.000
18	2	43.007	17.450	1.000
19	3	38.524	17.380	1.000

21	3	31.422	17.299	1.000
24	1	23.941	17.248	1.000
25	1	22.032	17.240	1.000
26	1	20.341	17.234	1.000
28	1	17.492	17.226	1.000
		a = 15468.0768 b = 2.0361 R^2 = 0.9953	a = 16.5427E+010 b = 5.1773 c = 8.3827 d = 17.2140 R^2 = 0.9995	a = 52145.7174 b = 0.8186 R^2 = 0.9994

As can be seen, all fittings are very good; the small differences between the R^2 values have no relevance. Yet, the Popescu function is the most realistic one with respect to the fact that extremely large polysemy values are very seldom while the other two functions strongly overestimate the tail of the data. The good fitting of the Mandelbrot function, which as the only one predicts adequate numbers in the middle part of the distribution, must be paid for with two additional parameters whose linguistic interpretation is not easy.

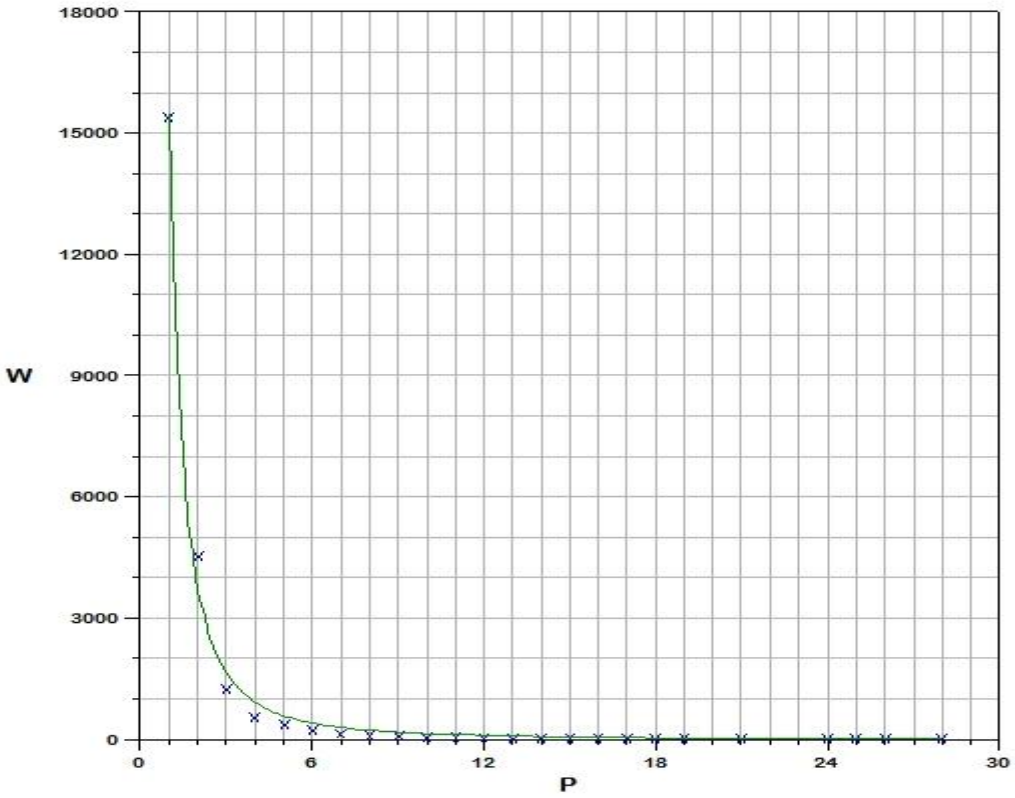

Figure 2: Fitting the power law function to the data

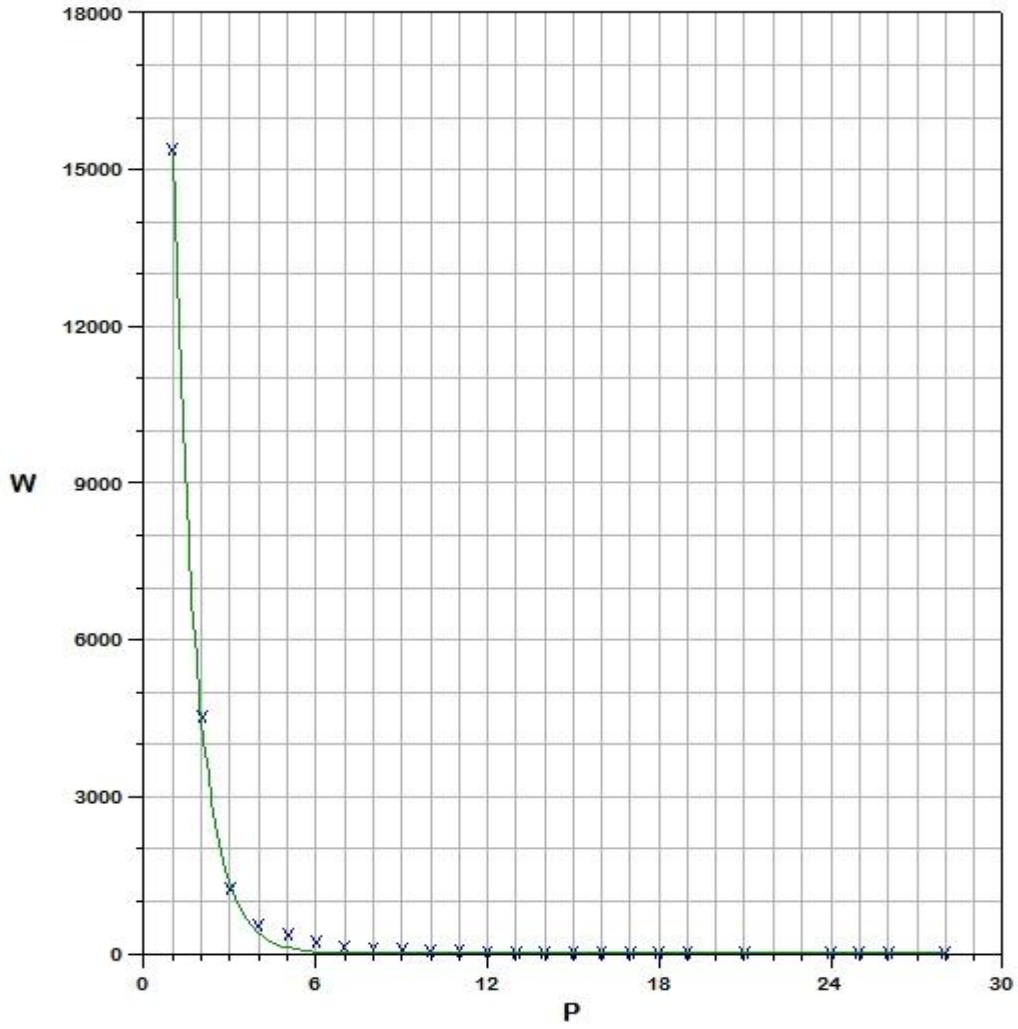

Figure 3. Fitting the Popescu function to the data

4.2. The polysemy and word-length relation

Polysemy and mean word length are calculated from our corpus data and shown in Table 4 and Figure 4, from which we can see that mean word length decreases with growing polysemy. Specifically, when polysemy grows up to 9, the mean word length reaches the minimum 1. Actually, all the words containing more than 8 meanings are one-character words and can also be found in Table 1.

Table 4
Polysemy values and the corresponding mean word lengths.

Polysemy	**Mean length**	**Power function (1)**
1	2.28805	2.416493
2	1.94967	1.610543
3	1.60325	1.373176
4	1.25581	1.263159
5	1.11582	1.200703
6	1.06220	1.160848
7	1.02439	1.133392
8	1.01176	1.113428
9	1	1.098313
10	1	1.086508
11	1	1.077055
12	1	1.069329
13	1	1.062909
14	1	1.057495
15	1	1.052875
16	1	1.048890
17	1	1.045421
18	1	1.042375
19	1	1.039683
21	1	1.035142
24	1	1.029883
25	1	1.028438
26	1	1.027115
28	1	1.024782
		a = 1.41649 b = 1.21416 R^2 = 0.8927

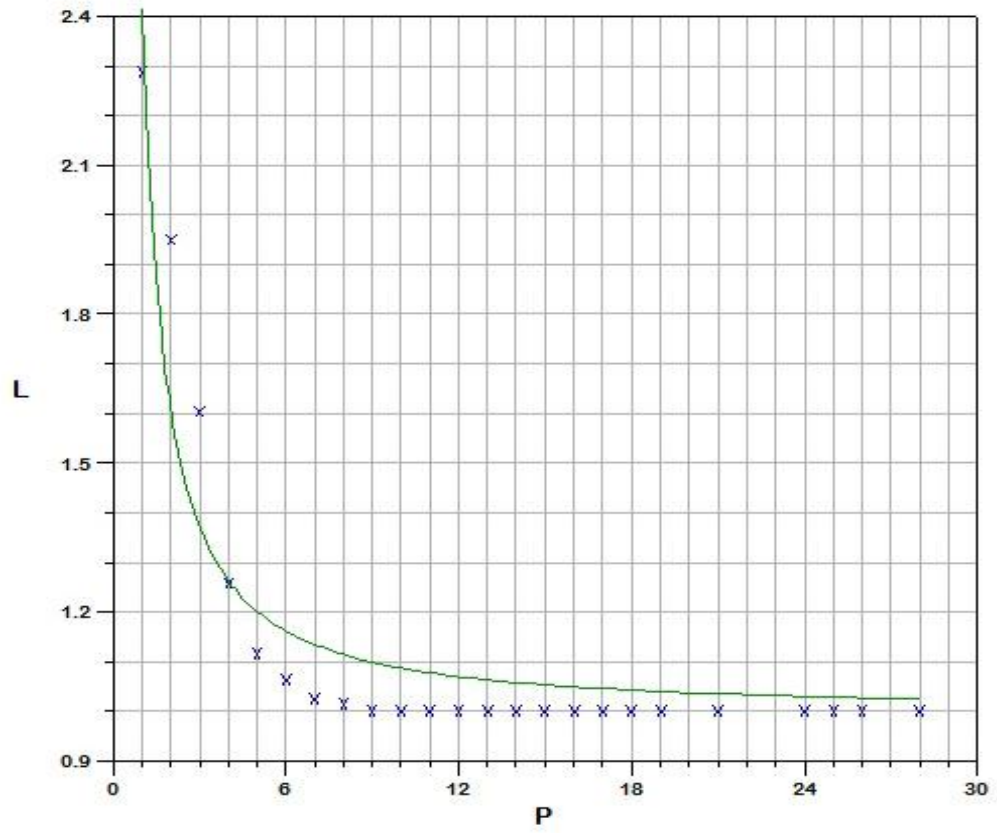

Figure 4. Word length as a function of polysemy ($L = 1 + aP^b$)

Table 5 shows the word length and mean polysemy values, which reveals that the longer a word, the less polysemy it includes.

Table 5
Word length and corresponding mean polysemy.

Word length	**Mean polysemy**	**Power function (2)**
1	4.51574	4.514657
2	1.33750	1.364606
3	1.15460	1.096869
4	1.04300	1.037824
5	1.04167	1.018238
6	1	1.010049
7	1	1.006071
		a = 3.51465729 b = 3.26897422 R^2 = 0.9995

Fitting a power law function (2) to the data in Table 5 yields $P = 3.5147\ L^{-3.2690}$ as shown in Figure 5. The coefficient of determination is $R^2 = 0.9995$, a results which indicates a slightly more significant fit than we obtained for the other direction; a finding which confirms with Köhler's report on his study on Maori (1999).

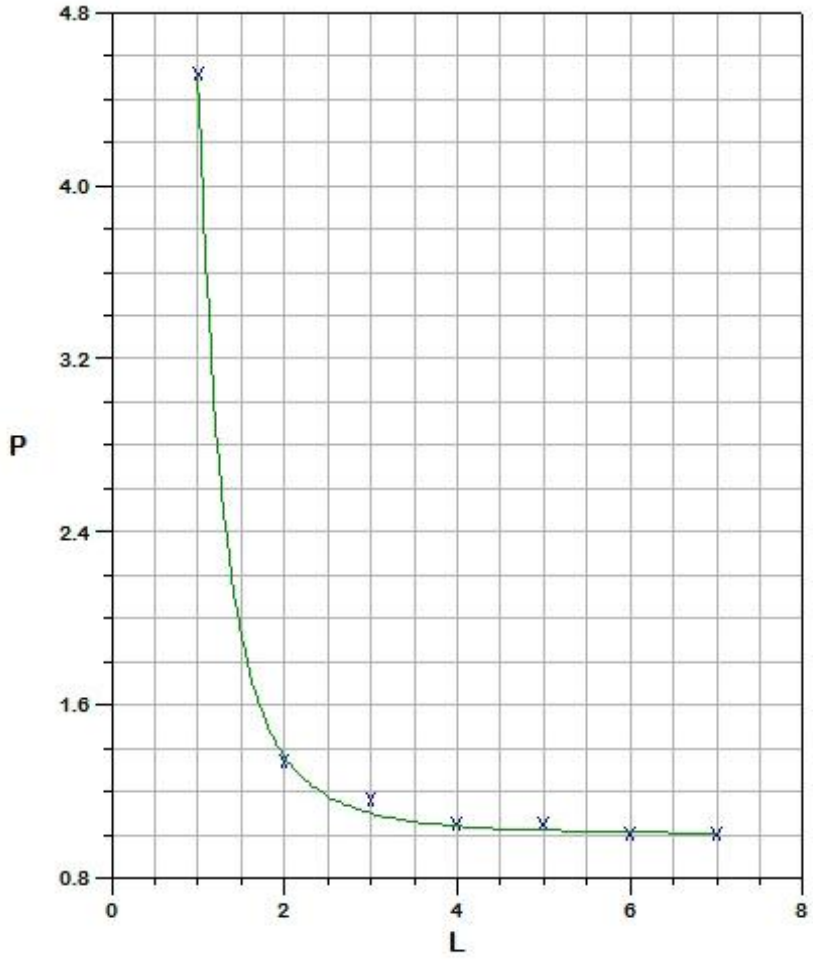

Figure 5. Polysemy as a function of word length ($P = 1 + aL^b$).

5. Conclusion

The Zipf-Mandelbrot, Zipf and Popescu hypothesis are successfully tested on dictionary data selected on the basis of a Chinese text corpus in the above analysis. This result is another support for the assumption that polysemy is lawfully distributed in all languages regardless of their typological characteristics.

Future investigations will have to scrutinize more and other data from Chinese in order to obtain a more general view of the quantitative properties of polysemy in this language.

References

Altmann, G., Beőthy, E., Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 537-543*

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107*.

Fickermann, I., Markner-Jäger, B, Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. *Glottometrika 6, 115-12*6.

Köhler, R. (1999). Der Zusammenhang zwischen Lexemlänge und Polysemie im Maori. In: Ondrejovič, S., Genzor, J. (eds.), *Pange lingua. Zborník na počest´ Viktora Krupu: 27-33*. Bratislava: Veda.

Popescu, I., Altmann, G., Köhler, R. (2010). Zipf's law – another view. *Quality and Quantity 44, doi: 10.1007/s11135-009-9234-y.*

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

Glottometrics 22, 2011, 85

Project

Quantitative Linguistics funded by NSSFC

Jin CONG, Zhejiang University

With the recent completion of 2011's annual selection of the key projects supported by the National Social Science Foundation of China (NSSFC), the project entitled "Quantitative Linguistic Research of Contemporary Chinese" (Grant No. 11&ZD188), with Professor Haitao Liu from Zhejiang University as the principal investigator, has been approved and initiated as one of the projects in the section of multidisciplinary research. The key projects supported by NSSFC, with the most substantial financial support granted and the highest authority attributed, are the top-ranked national government-funded projects at present in China's humanities and social sciences. This project headed by Haitao Liu has been the first quantitative linguistic key project funded by NSSFC.

The project consists of the following four sub-projects.

(1) Quantitative research of contemporary Chinese text. Through statistical analysis of samples from multiple genres of contemporary Chinese, it attempts to reveal the structural features of contemporary Chinese and the similarities and differences between its major genres.

(2) Quantitative research of contemporary Chinese syntax. With a syntactic formalism apt for the description of authentic Chinese text, syntactic annotation will be conducted of the corpus data collected. On the basis of the annotated data, the local features and global organization of contemporary Chinese, the local-global relationships, and the emergence of syntax will be addressed. For a deeper understanding of the various parameters and rules in the contemporary Chinese syntax, this sub-project will also collect and analyze a certain amount of diachronic language data of Chinese and synchronic cross-linguistic data, with an eye to the evolution of Chinese syntax and comparative quantitative studies of Chinese and other languages.

(3) Computation-oriented formalized quantitative research of natural language. The relationships between quantitative linguistics and computational linguistics will be investigated with the purpose of establishing, based on quantitative linguistic research, a system of computational linguistics guided by linguistic theories.

(4) Research of the relationships between quantitative studies of Chinese and cognitive structures. On the basis of large-scale authentic language data and through quantitative analysis and cognitive psychological experiments, it attempts to explore the relationships between the quantitative characteristics of language and the mechanisms of human cognition.

Sub-projects 1 and 2, which focus on quantitative description and comparison, constitute the primary task of this project. These two lines of inquiry will pave the way for sub-projects 3 and 4, which are devoted to the bridge between linguistics and other sciences. As methodological innovation is always a vital part in the development of quantitative linguistics, it is what this project practices throughout the fulfillment of its goals.