Glottometrics 21, 2011, 54-59

On stratification in poetry

Ioan-Iovitz Popescu, Bucharest
Radek Čech, Ostrava
Gabriel Altmann, Lüdenscheid

Abstract. Texts are composed of many different strata on different levels. A method is proposed to find the number of strata at the word-form level in Slovak poetry and to study the relationship between the parameters of the fitting function.

Keywords: Rank-frequency distribution, word forms, stratification, Slovak poetry

Strata arise in texts by mixing different means of expression which can be of quite variegated kind. For example words from different word classes, words of different length, interjections, different sentence types, different pictures of reality, etc., may bring about a kind of stratification. The methods of investigation are very few; as a matter of fact, there is only one work in which the rank-frequency distribution of word-forms – the so called "Zip's law" – is considered a result of stratification (cf. Popescu, Altmann, Köhler 2010).

If we knew what classes are present in a writer's brain at the moment of writing, we would be able to separate them. However, this is not possible in general and every step in this direction is merely a trial, an empirical approximation of the state of the affairs. If one supposes the existence of stratification, one may scrutinize the phenomenon by setting up the rank-frequency distribution of some supposed classes. The rank-frequency distribution of linguistic entities abides by the function

$$(1) \qquad y = 1 + a \bullet \exp(-x/b) + c \bullet \exp(-x/d) + \ldots..$$

The number of exponential components signalizes the number of strata. The constant 1 is added because frequencies cannot be smaller, hence the function converges to 1. In difference to polynomials whose use in text analysis cannot be recommended, the above function shows which components are redundant: if the constants in the exponents of two components are equal or almost equal, then one of the components is redundant and can be omitted.

In order to illustrate this property we computed the rank-frequency distribution of word forms in the poem *Aby spriesvitnela* by Eva Bachletová and obtained the result in Table 1

Table 1
Rank-frequency distribution of word forms
in Bachletová's poem *Aby spriesvitnela*

r	f_r	r	f_r	r	f_r	r	f_r
1	4	14	1	27	1	40	1
2	3	15	1	28	1	41	1
3	3	16	1	29	1	42	1
4	2	17	1	30	1	43	1
5	2	18	1	31	1	44	1
6	2	19	1	32	1	45	1
7	1	20	1	33	1	46	1
8	1	21	1	34	1	47	1
9	1	22	1	35	1	48	1
10	1	23	1	36	1	49	1
11	1	24	1	37	1	50	1
12	1	25	1	38	1	51	1
13	1	26	1	39	1	52	1
						53	1

Fitting formula (1) to these data using only one component we obtain

$$f_r = 1 + 4.2851 \exp(-r/2.9793)$$

yielding the determination coefficient $R^2 = 0.955$. If we add a second component, we obtain

$$f_r = 1 + 1.9208 \exp(-r/2.9793) + 2.3823 \exp(-r/2.9793)$$

with the same $R^2 = 0.955$. As can be seen, the constants in the exponents are equal and the sum of the multiplicative constants yields approximately the amplitude in the one-component expression.

Hence we can conclude that the given poem is monostratal. Whatever force controls the word-form strata, it is not sufficiently expressed.

The rank-frequency distribution in the given poem and the fitting function are presented in Figure 1.

In order to study this property, we performed the same fitting in 54 poems of the same author and tested whether one component in (1) is sufficient. The results are

presented in Table 2. As can be seen, all of the poems are non-stratified and express a special feature of author style. Some comments to the results are in order here.

(1) In two cases the determination coefficient is smaller than 0.8 but testing the parameters and the regression by *t*- and *F*-tests yielded always highly significant results ($P < 0.0001$). Thus in all cases the theory of background stratification can be considered as corroborated by these data. If we compare the present fitting with the traditional "Zipfian" one using the power function, we can see that in each case function (1) yields better results.

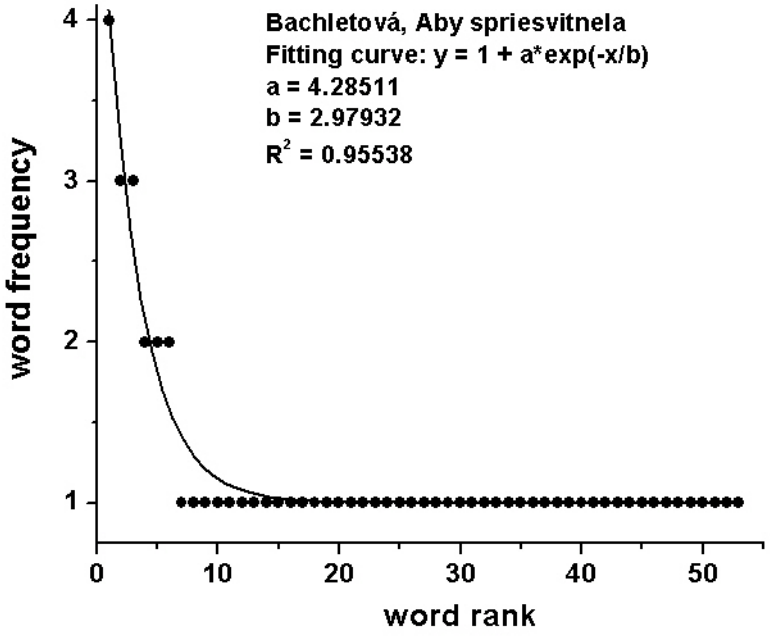

Figure 1. Rank-frequency distribution in E. Bachletová's poem

(2) Some values of parameter *a* seem to attain senseless magnitudes. This is caused by the fact that only one word has been repeated two or three times; all the rest of words occur only once. On the other hand, in all those cases the parameter *b* is very small, the exponential itself converges quickly to zero and only the constant 1 is relevant. Poems of this kind can be omitted from some kinds of analysis, because the word-form distribution is almost uniform, i.e. it does not display any tendency.

Table 2
Parameters of the first stratum in poems by E.Bachletová

Poem	**a**	**b**	**R^2**
Aby spriesvitnela	4.2851	2.9793	0.96
Bez rozlúčky	1.8587	2.2925	0.81
Čakáme šťastie...	3.9093	1.8886	0.89
Čakanie na Boží jas	15.2546	1.4538	0.96
Čas pre nádych vône	3.218	3.8553	0.92

Dielo Stvoriteľa	9.1363	2.4957	0.9
Dnešný luxus	4.7167	2.2445	0.96
Do večnosti beží čas	4.7908	2.9642	0.95
Dovoľ mi slúžiť	6586.9417	0.1137	1
Ešte raz	4.5897	2.5977	0.95
Hľadanie odpovedí	2.1771	4.9942	0.86
Iba neha	13.5437	2.6483	0.84
Iba život	12641.6345	0.1143	1
Idem za Tebou	2.4715	3.5407	0.88
Ihly na nebi	3.7075	4.8980	0.92
Istota	4.7167	2.2445	0.96
Keď dohorí deň	6.8233	1.7770	0.97
Kým ich máme	7.0289	1.1594	0.97
Len áno	1.5774	5.5358	0.75
Malé modlitby	2.1771	4.9943	0.85
Malý ošiaľ	5.5359	4.5479	0.85
Miesto pre Nádej	6586.9417	0.1137	1
Moje určenie	24.1348	1.1031	0.87
Nado mnou Ty sám...	10.7635	0.7873	0.99
Náš chrám	24.6841	0.8768	0.96
Naše mamy	5.2926	1.8678	0.97
Naše svetlo	5.9052	4.2045	0.95
Neopusť ma...	10.9337	1.2572	0.99
Nepoznateľné	8.8198	2.9232	0.94
Podobnosť bytia	46.1763	0.5690	0.95
Pravidlá odpúšťania	4.5000	1.4427	0.83
Precitnutie	3.1155	2.1840	0.92
Prvotný sen	12.6744	1.0789	0.99
Rozdelená bytosť	2.1771	4.9942	0.86
Rozťatá prítomnosť	3.3933	6.1867	0.93
Som iná	10.3289	1.7111	0.94
Spájania	3.9093	1.8886	0.89
Stály smútok pre šesť písmen	12.0732	4.1606	0.93
Tá Láska	1.6499	3.9236	0.78
Tak málo úsmevu	38.1267	0.5887	0.99
Ťažko pokoriteľní	3.8321	1.5666	0.94
Tiché verše	2.2500	1.4427	0.83
To všetko je dar	6.4480	3.9576	0.89
Ulomené zo slov	2.7206	2.8457	0.89
Vďaka Pane!	2.2500	1.4427	0.83
Vďaka za deň	1.8587	2.2925	0.81
Večerná ruža	12641.6345	0.1143	1
Večerné ticho	6.1350	0.2421	0.92

Vo večnosti slobodná	9.0387	4.8815	0.96
Vrátili sa	3.1155	2.1840	0.92
Vyznania	2.7206	2.8457	0.9
Z neba do neba	7.6899	2.0037	0.96
Zasľúbenie jasu	2.8415	4.4980	0.82
Zbytočné srdce	13.4423	0.9778	0.92

(3). Omitting the poems with abnormal parameter *a* we can easily state the relationship between the parameters *a* and *b*. In general, one expects a monotonously decreasing $a = f(b)$ because parameter *a* is merely a balancing magnitude responsible for the amplitude of (1). The decrease of frequencies is controlled by parameter *b.* Thus ordered according to increasing *b* we obtain the values presented in Table 3.

Table 3
Relationship between parameters a and b

b	**a**	$\mathbf{a_{theor}}$	**b**	**a**	$\mathbf{a_{theor}}$
0.5690	46.1763	43.3451	2.2925	1.8587	5.2659
0.5887	38.1267	40.0943	2.4957	9.1363	5.0456
0.7873	10.7635	21.3054	2.5977	4.5897	4.9571
0.8768	24.6841	17.1926	2.6483	13.5437	4.9176
0.9778	13.4423	14.0305	2.8457	2.7206	4.7863
1.0789	12.6744	11.8396	2.8457	2.7206	4.7863
1.1031	24.1348	11.4169	2.9232	8.8198	4.7431
1.1594	7.0289	10.5508	2.9642	4.7908	4.7218
1.2572	10.9337	9.3563	2.9793	4.2851	4.7142
1.4427	4.5000	7.8136	3.0062	6.1350	4.7011
1.4427	2.2500	7.8136	3.5407	2.4715	4.5073
1.4427	2.2500	7.8136	3.8553	3.218	4.4342
1.4538	15.2546	7.7426	3.9236	1.6499	4.4210
1.5666	3.8321	7.1178	3.9576	6.4480	4.4147
1.7111	10.3289	6.5177	4.1606	12.0732	4.3809
1.7770	6.8233	6.2991	4.2045	5.9052	4.3743
1.8678	5.2926	6.0412	4.4980	2.8415	4.3361
1.8886	3.9093	5.9882	4.5479	5.5359	4.3304
1.8886	3.9093	5.9882	4.8815	9.0387	4.2977
2.0037	7.6899	5.7289	4.8980	3.7075	4.2963
2.1840	3.1155	5.4147	4.9942	2.1771	4.2883
2.1840	3.1155	5.4147	4.9942	2.1771	4.2883
2.2445	4.7167	5.3286	4.9943	2.1771	4.2883
2.2445	4.7167	5.3286	5.5358	1.5774	4.2522
2.2925	1.8587	5.2659	6.1867	3.3933	4.2226

The given relationship can be represented by a simple power function

$$a = 4.1317 + 9.3496b^{-2.5426}$$

yielding an $R^2 = 0.79$ and very highly significant *t*- and *F*-values. It can be expected that adding further poems by the same author the relationship will get rather stronger. The relationship is graphically presented in Figure 2.

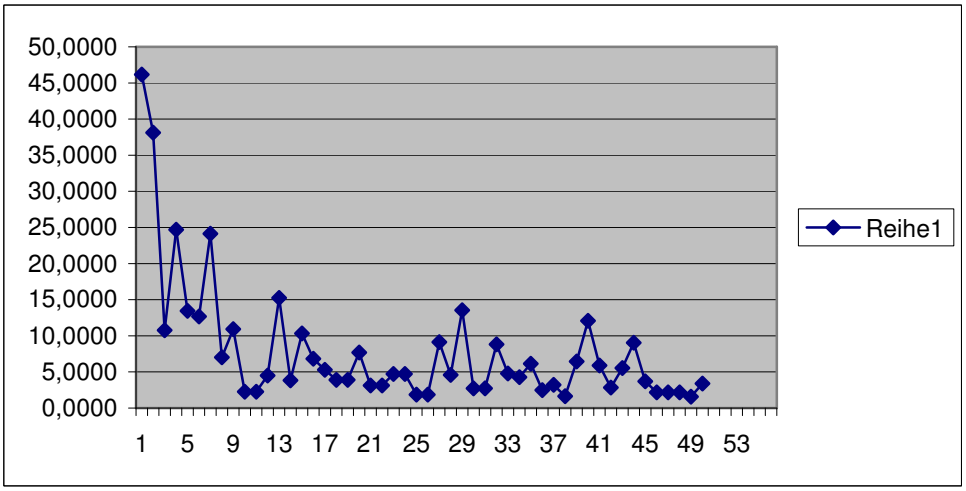

Figure 2. The relationship between parameters *a* and *b*

(4) Automatically, further questions arise that can be pursued in the future: (a) Does the above result hold only for the given writer or can we transfer it to other writers, too? (b) Does it hold only for poetry of this kind (without rhyme, irregular verse) or does it hold for Slovak poetry in general? (c) Does it hold also for poetry in other languages? (d) Does it hold also for prose?

In short texts, the lemmatization of the word forms does not bring new results, not even in strongly synthetic languages. In strongly analytic ones the results are almost identical.

The above result is a strong support for replacing the Zipfian zeta function by formula (1).

References

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf´s law – another view. *Quality and Quantity 44(4), 713-731.*

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