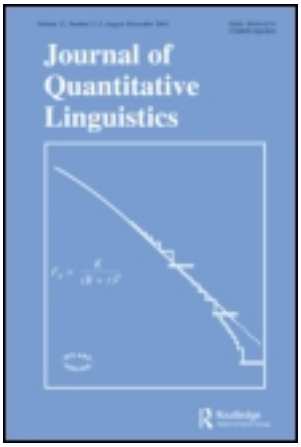

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/njql20

Some Geometric Properties of Slovak Poetry*

Ioan-Iovitz Popescu [a] , Radek Čech [a] & Gabriel Altmann [a]
[a] The Romanian Academy, Bucharest, Romania
[b] University of Ostrava, Ostrava, Czech Republic
[c] Ruhr-Universität Bochum, Bochum, Germany

Available online: 13 Mar 2012

To cite this article: Ioan-Iovitz Popescu, Radek Čech & Gabriel Altmann (2012): Some Geometric Properties of Slovak Poetry*, Journal of Quantitative Linguistics, 19:2, 121-131

To link to this article: http://dx.doi.org/10.1080/09296174.2012.659000

Journal of Quantitative Linguistics
2012, Volume 19, Number 2, pp. 121–131
http://dx.doi.org/10.1080/09296174.2012.659000

Routledge
Taylor & Francis Group

Some Geometric Properties of Slovak Poetry*

Ioan-Iovitz Popescu[1], Radek Čech[2] and Gabriel Altmann[3]
[1]The Romanian Academy, Bucharest, Romania; [2]University of Ostrava, Ostrava, Czech Republic; [3]Ruhr-Universität Bochum, Bochum, Germany

ABSTRACT

The article is focused on the analysis of the frequency structure of texts. Specifically, a geometric characterization of the rank-frequency sequence, which is determined by relationships among the highest word frequency, a number of particular word forms in a text, and the so-called *h*-point, is analysed. We observe that the geometric characterization of the rank-frequency sequence correlates strongly with the other frequency-based text property called the golden section of text which seems to be a result of a universal mechanism controlling one's language usage. These findings can be viewed as a result of self-organization of text.

1. INTRODUCTION

In several previous works (cf. Popescu & Altmann, 2006, 2007; Popescu et al., 2009, pp. 73ff) it has been shown that the *h*-point whose location in the rank-frequency sequence or in the frequency spectrum of word forms can be exactly specified, may serve as a characteristic of text or of style. (In sociometrics it is called Hirsch's index or h-index (Hirsch 2005); in linguistics it has been introduced by Popescu 2007). From the creative point of view, one can imagine the writer positioned at the *h*-point and observing the development of word repetitions above the *h*-point and the increase of types below the *h*-point. The *h*-point is a fixed point separating auto semantics occurring seldom from the synsemantics occurring more frequently. Those auto semantics whose frequency is greater than the *h*-point belong to the thematic domain and the *h*-point helps to identify them.

*Address correspondence to: Ioan-Iovitz Popescu. E-mail: iovitzu@gmail.com

0929-6174/12/19020121

In this paper we want to study the geometric aspects of the *h*-point as applied to the works of a unique writer. We selected Bachletová, a Slovak poet and her 54 poems which were placed at our disposal by the writer herself but can be found also on the Internet (http://www.evabachletova.sk). We shall study only the rank-frequency data and leave the frequency spectrum aside because it usually contains very few frequency classes.

2. A GEOMETRIC CHARACTERIZATION OF THE RANK-FREQUENCY SEQUENCE

The *h*-point has been defined in the literature (Popescu et al., 2009, p. 17) as

$$h = \begin{cases} r, & \text{if there is an } r = f(r) \\ \frac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \text{if there is no } r = f(r) \end{cases} \tag{1}$$

and can be given simply as that rank r which is equal to its frequency $f(r)$. If there is no such r, one solves the second part of (1) and obtains a positive real number which can be left in its form or rounded or truncated.

Besides the *h*-point there are two other prominent points in the sequence, namely rank 1 and its frequency, i.e. $(1, f(1))$, and the greatest rank which is equal to the vocabulary V and has the frequency 1, i.e. $(V, 1)$. Only in extremely long texts is the smallest frequency greater than 1 – a very rare case, but in such cases one simply takes $(V, f(V))$. Since the rank-frequency sequence mostly has a hyperbolic form, one can easily imagine the triangle defined by the points $P_1(V, f(V))$, $P_2(1, f(1))$, $P_3(h, h)$, and presented in Figure 1. As a matter of fact, this triangle is an approximation to the area between the points P_1 and P_2 and the frequency sequence. Instead of using a sum of trapezoids we simplify the computation by this approximation.

The area of the triangle between the three conspicuous points can be computed as

$$A_h = \frac{1}{2}\begin{vmatrix} V & 1 & 1 \\ 1 & f(1) & 1 \\ h & h & 1 \end{vmatrix} \tag{2}$$

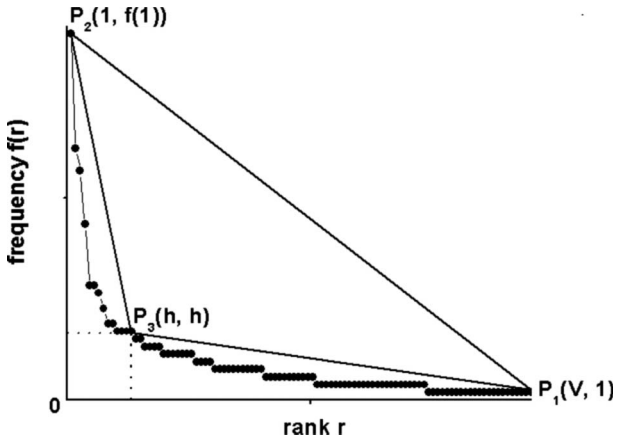

Fig. 1. The characteristic triangle of the rank frequency sequence (from Popescu & Altmann, 2006).

where $f(V)$ (second element in the first row) has been replaced by $f(1)$. If the greatest rank does not have frequency 1, all frequencies can be transformed in such a way that the last rank has frequency 1 namely by $f(r') = f(r) - f(V) + 1$. The evaluation of (2) yields

$$A_h = \frac{1}{2}|Vf(1) + 2h - h(V + f(1)) - 1| \tag{3}$$

This indicator can be relativized in order to be comparable. First we compute $A_{\max}$ which can be obtained by replacing $P_3(h, h)$ by $P_3(1, 1)$ leading to

$$A_{\max} = (1/2)(V - 1)(f(1) - 1) \tag{4}$$

Then the ratio

$$A = A_h / A_{\max} \tag{5}$$

shows the extent of exploitation of the given vocabulary. If A is small then many words must be used frequently. The greater A is, the more words have very small frequency, i.e. only few words are used often, the rest are rare words, mostly autosemantics. As can be seen, A depends only on V and $f(1)$; text length N is not relevant, though further examinations are necessary (cf. Popescu et al., 2009, p. 80) taking into

account many long and many very short texts. However, if the sequence is not strictly hyperbolic, e.g. if the h-point is equal to $f(1)$ or if $f(1) = 1 - a$ case that may occur with very short texts – the line $\overline{P_1P_2}$ lies below the h-point and the triangle is very small, we obtain a small A. If $f(1) = 1$, then we set by definition $A = 0$.

For the sake of illustration we show the computation of A using a short text *Bez rozlúčky* by Bachletová as presented in Table 1. Here $h = 2$, hence A_h computed by means of (3) yields for $f(1) = 2$, $h = 2$, $V = 32$

$$A_h = (1/2)|32(2) + 2(2) - 2(32 + 2) - 1| = 0.5$$

$$A_{\max} = (1/2)(32 - 1)(2 - 1) = 15.5$$

hence

$$A = 0.5/15.5 = 0.0323.$$

Let us look at the other poems by Bachletová as presented in Table 2.

It would be premature to set up hypotheses about the distribution of A because there is little data (54 poems). With Bachletová, $A \in < 0.0286$;

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

Rank r	Word form	Frequency $f(r)$	Rank r	Word form	Frequency $f(r)$
1	a	2	17	nej	1
2	bez	2	18	odišla	1
3	na	2	19	predzvesťou	1
4	čakajú	1	20	rozlúčky	1
5	deti	1	21	s	1
6	díva	1	22	sa	1
7	do	1	23	sveta	1
8	drahá	1	24	svoje	1
9	ja	1	25	ticha	1
10	kde	1	26	tu	1
11	mama	1	27	večnosti	1
12	mamy	1	28	viem	1
13	mňa	1	29	všetky	1
14	moja	1	30	z	1
15	náhle	1	31	že	1
16	nebies	1	32	žijem	1

Table 2. Computation of A in poems by Bachletová.

	N	V	h	$f(1)$	A_h	A_{max}	A
Aby spriesvitnela	63	53	3	4	23.00	78.0	0.2949
Bez rozlúčky	35	32	2	2	0.50	15.5	0.0323
Čakáme šťastie...	48	43	2.5	3	9.00	42.0	0.2143
Čakanie na Boží jas	77	59	3	9	166.00	232.0	0.7155
Čas pre nádych vône	81	74	2.5	3	16.75	73.0	0.2295
Dielo Stvoriteľa	136	113	3	8	273.00	392.0	0.6964
Dnešný luxus	36	28	2.5	4	18.00	40.5	0.4444
Do večnosti beží čas	51	40	3	4	16.50	58.5	0.2821
Dovoľ mi slúžiť	34	33	1.5	2	7.75	16.0	0.4844
Ešte raz	42	33	3	4	13.00	48.0	0.2708
Hľadanie odpovedí	67	58	2	3	27.50	57.0	0.4825
Iba neha	139	93	3.67	12	368.66	506.0	0.7286
Iba život	44	42	1.33	3	33.83	41.0	0.8252
Idem za Tebou	72	65	2	3	31.00	64.0	0.4844
Ihly na nebi	54	39	3	4	16.00	57.0	0.2807
Istota	41	33	2.5	4	21.75	48.0	0.4531
Keď dohorí deň	52	43	2.5	5	49.50	84.0	0.5893
Kým ich máme	44	39	2	4	36.50	57.0	0.6404
Len áno	34	27	2	2	−0.50	13.0	0.0385
Malé modlitby	51	42	2	3	19.50	41.0	0.4756
Malý ošiaľ	68	45	3	7	82.00	132.0	0.6212
Miesto pre Nádej	29	28	1.5	2	6.50	13.5	0.4815
Moje určenie	146	115	3	11	446.00	570.0	0.7825
Nado mnou Ty sám...	38	34	2	4	31.50	49.5	0.6364
Náš chrám	86	72	2.5	9	224.75	284.0	0.7914
Naše mamy	56	49	2.5	4	33.75	72.0	0.4688
Naše svetlo	67	47	4	5	17.00	92.0	0.1848
Neopusť ma...	36	27	2.5	6	41.75	65.0	0.6423
Nepoznateľné	93	69	3	8	163.00	238.0	0.6849
Podobnosť bytia	85	72	2	9	244.50	284.0	0.8609
Pravidlá odpúšťania	35	31	2.33	3	8.67	30.0	0.2889
Precitnutie	51	46	2	3	21.50	45.0	0.4778
Prvotný sen	81	73	2.5	6	122.25	180.0	0.6792
Rozdelená bytosť	79	70	2	3	33.50	69.0	0.4855
Rozťatá prítomnosť	78	60	3	4	26.50	88.5	0.2994
Som iná	58	43	2.67	7	86.00	126.0	0.6825
Spájania	48	43	2.5	3	9.00	42.0	0.2143
Stály smútok pre šesť písmen	146	93	4	12	351.50	506.0	0.6947
Tá Láska	35	30	2	2	0.50	14.5	0.0345
Tak málo úsmevu	63	54	2	8	155.50	185.5	0.8383
Ťažko pokoriteľní	30	26	2	3	11.50	25.0	0.4600

(*continued*)

Table 2. (*Continued*).

	N	V	h	$f(1)$	A_h	A_{max}	A
Tiché verše	31	29	2	2	0.50	14.0	0.0357
To všetko je dar	58	38	4.33	5	5.67	74.0	0.0766
Ulomené zo slov	31	25	2	3	11.00	24.0	0.4583
Vďaka Pane!	37	35	2	2	0.50	17.0	0.0294
Vďaka za deň	39	36	2	2	0.50	17.5	0.0286
Večerná ruža	46	44	1.67	3	28.00	43.0	0.6511
Večerné ticho	68	53	3	6	73.00	130.0	0.5615
Vo večnosti slobodná	170	126	4	9	300.50	500.0	0.6010
Vrátili sa	51	46	2	3	21.50	45.0	0.4778
Vyznania	55	49	2	3	23.00	48.0	0.4792
Z neba do neba	67	55	3	6	76.00	135.0	0.5630
Zasľúbenie jasu	52	41	2	4	38.50	60.0	0.6417
Zbytočné srdce	36	27	2	6	49.50	65.0	0.7615

$0.8609>$ and the mean is 0.4752. No distribution trend is visible and it can be shown that A does not correlate with other measures of vocabulary richness. It seems to be simply a geometric characterization of the rank-frequency sequence. However, we shall show in Section 3 that it is a result of a kind of self-organization.

3. THE GOLDEN SECTION

Writing a text, its author subconsciously observes the repetition of words, cares for good style (usually avoiding unacceptable repetitions) and tries to keep balance between auxiliaries and autosemantics. In poetry the situation is slightly different, because auxiliaries are usually short words that fill space, help to control the rhythm, etc. Writing is an extremely complex process full of mental and grammatical linearizations, corrections, pauses, etc. The longer the text, the more the writer loses his subconscious control over some proportions and keeps conscious control only over contents, grammar, his aim, etc. But as soon as parts of his control disappear, the text develops its own dynamics and begins to abide by some laws which are not known to the writer but work steadily in the background. The process is analogous to that in physics: if we walk, we consider our activity as something normal; but if we stumble, i.e. lose the control, gravitation manifests its presence and we fall. That means, gravitation does not work ad hoc in order to worry us maliciously,

but it is always present, even if we do not realize it consciously. In writing, laws are present, too, and they work at a level which is only partially accessible. One can overcome their working, but one cannot eliminate them. On the other hand, if the writer slowly loses his control of frequency structuring, a new order begins to arise by self-organization or by some not perceivable background mechanism.

As has been shown in other publications (cf. Popescu & Altmann, 2007; Tuzzi et al., 2010b), one such background mechanism is the striving for the golden section in texts. The writer does not strive for it and does not even know that it exists or, if so, how to attain it. Much pains have been taken with its discovery at all (Livio, 2002; Doczi, 2005; Hemenway, 2005; Sahlquist, 2008) but its existence is known in all arts. In writing, the tendency to approximate the golden ratio has been shown both in short texts (e.g. sequentially in Goethe's *Erlkönig*), in very long texts in 20 languages, and in 60 New Year Speeches of Italian presidents (cf. Tuzzi et al., 2010a). Here we shall try to capture the tendency in 54 poems by Bachletová.

If we look from the h-point at the two extreme points of the rank-frequency sequence $P_1(V, 1)$ and $P_2(1, f(1))$, then the angle α at the crossing of P_3P_1 with P_3P_2 can be called the writer's view. "Sitting" at this point the writer so to say controls the harmony in the development of autosemantics and auxiliaries. In short texts this is possible at least subconsciously, but the longer the text becomes, the more it develops its own dynamics and controls the writer himself (to quote the famous dictum of Macchiavelli). The proportions of words become obscured and the writing becomes more spontaneous. This dynamics causes that the radian of the angle α strives towards the golden ratio given as

$$\varphi = \frac{1+\sqrt{5}}{2} = 1.6180,$$

which, in this case, represents a kind of fixed point attractor. For our purposes we can compute the exact value of $cos\ \alpha$ as

$$cos\,\alpha = \frac{-[(h-1)(f(1)-h)+(h-1)(V-h)]}{[(h-1)^2+(f(1)-h)^2]^{1/2}[(h-1)^2+(V-h)^2]^{1/2}} \qquad (6)$$

and show that α expressed in radians tends to the golden section. The greater the control of the writer, the greater is the distance of α rad from

the golden section. For the works by Bachletová we present the results in Table 3.

As can be seen, the control with Bachletová is strong. Nevertheless, it can be shown that with increasing length of the poem it is directed towards the golden section, as is shown in Figure 2.

The trend may be considered linear but the dispersion is too great to yield a straight line.

Table 3. "Writer's view" in poems by Bachletová.

Poem	*N*	*V*	*h*	*f*(1)	*cos* α	α *rad*
Aby spriesvitnela	63	53	3	4	−0.9116	2.7179
Bez rozlúčky	35	32	2	2	−0.9994	3.1083
Čakáme šťastie...	48	43	2.5	3	−0.9597	2.8569
Čakanie na Boží jas	77	59	3	9	−0.3499	1.9282
Čas pre nádych vône	81	74	2.5	3	−0.9551	2.8408
Dielo Stvoriteľa	136	113	3	8	−0.3882	1.9695
Dnešný luxus	36	28	2.5	4	−0.7474	2.4150
Do večnosti beží čas	51	40	3	4	−0.9173	2.7319
Dovoľ mi slúžiť	34	33	1.5	2	−0.7182	2.3721
Ešte raz	42	33	3	4	−0.9222	2.7445
Hľadanie odpovedí	67	58	2	3	−0.7196	2.3740
Iba neha	139	93	3.67	12	−0.3331	1.9103
Iba život	44	42	1.33	3	−0.2041	1.7764
Idem za Tebou	72	65	2	3	−0.7182	2.3721
Ihly na nebi	54	39	3	4	−0.9179	2.7334
Istota	41	33	2.5	4	−0.7410	2.4053
Keď dohorí deň	52	43	2.5	5	−0.5459	2.1482
Kým ich máme	44	39	2	4	−0.4712	2.0615
Len áno	34	27	2	2	−0.9992	3.1016
Malé modlitby	51	42	2	3	−0.7246	2.3812
Malý ošiaľ	68	45	3	7	−0.4893	2.0820
Miesto pre Nádej	29	28	1.5	2	−0.7203	2.3751
Moje určenie	146	115	3	11	−0.2598	1.8336
Nado mnou Ty sám...	38	34	2	4	−0.4749	2.0657
Náš chrám	86	72	2.5	9	−0.2458	1.8192
Naše mamy	56	49	2.5	4	−0.7295	2.3884
Naše svetlo	67	47	4	5	−0.9684	2.8895
Neopusť ma...	36	27	2.5	6	−0.4494	2.0368
Nepoznateľné	93	69	3	8	−0.3993	1.9816
Podobnosť bytia	85	72	2	9	−0.1555	1.7270
Pravidlá odpúšťania	35	31	2.33	3	−0.9142	2.7244
Precitnutie	51	46	2	3	−0.7230	2.3789
Prvotný sen	81	73	2.5	6	−0.4134	1.9970

(*continued*)

Table 3. (*Continued*).

Poem	*N*	*V*	*h*	*f*(1)	*cos* α	α *rad*
Rozdelená bytosť	79	70	2	3	−0.7174	2.3709
Rozťatá prítomnosť	78	60	3	4	−0.9096	2.7130
Som iná	58	43	2.67	7	−0.3972	1.9793
Spájania	48	43	2.5	3	−0.9597	2.8569
Stály smútok pre šesť písmen	146	93	4	12	−0.3825	1.9633
Tá Láska	35	30	2	2	−0.9994	3.1059
Tak málo úsmevu	63	54	2	8	−0.1833	1.7552
Ťažko pokoriteľní	30	26	2	3	−0.7359	2.3978
Tiché verše	31	29	2	2	−0.9993	3.1046
To všetko je dar	58	38	4.33	5	−0.9951	3.0429
Ulomené zo slov	31	25	2	3	−0.7372	2.3996
Vďaka Pane!	37	35	2	2	−0.9995	3.1113
Vďaka za deň	39	36	2	2	−0.9996	3.1122
Večerná ruža	46	44	1.67	3	−0.4613	2.0502
Večerné ticho	68	53	3	6	−0.5875	2.1988
Vo večnosti slobodná	170	126	4	9	−0.5354	2.1358
Vrátili sa	51	46	2	3	−0.7230	2.3789
Vyznania	55	49	2	3	−0.7220	2.3775
Z neba do neba	67	55	3	6	−0.5863	2.1972
Zasľúbenie jasu	52	41	2	4	−0.4700	2.0601
Zbytočné srdce	36	27	2	6	−0.2811	1.8558

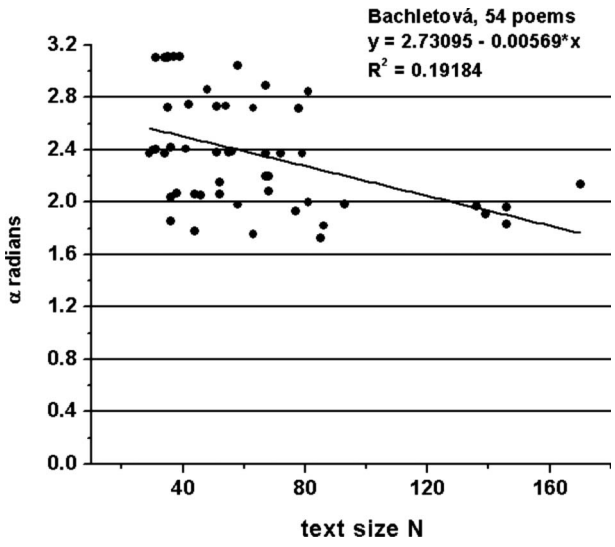

Fig. 2. The change of α rad with increasing text length with Bachletová.

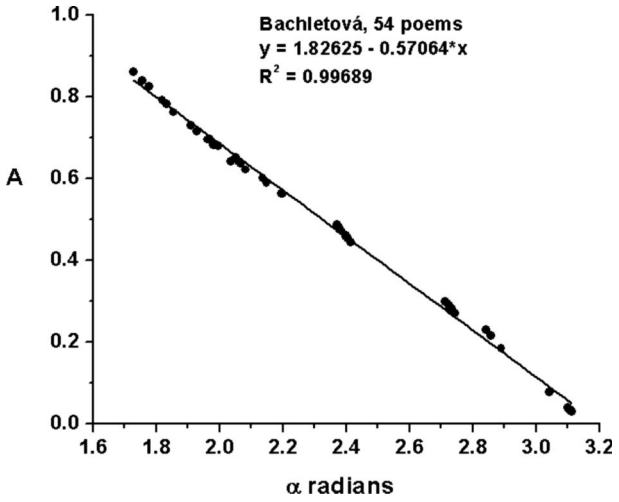

Fig. 3. The linear relationship $<\alpha$ rad. A$>$.

Since α rad converges to the golden section, we may in turn ask whether A is correlated with it. If so, then there would be a kind of order controlling the geometric frequency structuring of text. As a matter of fact, it can be shown that

$$A = 1.8263 - 0.5706\ (\alpha \text{ rad})$$

with $R^2 = 0.997$, i.e. there is an almost perfect linear relationship as can be seen in Figure 3.

4. CONCLUSIONS

The computation of A and α rad furnishes us with some characteristic features of the frequency structure of texts. At the same time, we see a rise of a kind of self-organization far from equilibrium. At the beginning of writing, the writer has strong control of the frequency structure; he "sees" the proportions of his text. But with increasing text he loses control and the frequency structure performs self-organization in that it strives to the golden section.

There are two tasks that should be performed in the future. (1) It is necessary to find other characteristic features with which A and α rad are

associated. Further, their properties (formal or semantic) have to be observed. As is well known, there are no isolated properties in language; hence this would be one of the possible continuations. (2) Perhaps the meaning of these two indicators would become clearer if we observed them in an increasing text, say taking 50 words, then 100, etc. In Table 3 and Figure 2 we have a mixture of texts and only the final state in each. Does it develop linearly or hyperbolically? And trying to answer the question ''why?'' we would make a step towards theory.

ACKNOWLEDGEMENTS

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

REFERENCES

Doczi, G. (2005). *The Power of Limits: Proportional Harmonies in Nature, Art and Architecture*. Boston: Shambhala Publications.

Hemenway, P. (2005). *Divine Proportion; Phi in Art, Nature, and Science*. New York: Sterling.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America 102, (46)*, 16569–16572.

Livio, M. (2002). *The Golden Ratio: The Story of Pi. The World's Most Astonishing Number*. New York: Broadway Books.

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 (Quantitative Linguistics)*. (pp. 557–567). Berlin/New York: Mouton de Gruyter.

Popescu, I.-I., & Altmann, G. (2006). Some geometric properties of word frequency distributions. *Göttinger Beiträge zur Sprachwissenschaft, 13*, 87–98.

Popescu, I.-I., & Altmann, G. (2007). Writer's view of text generation. *Glottometrics, 15*, 71–81.

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

Sahlquist, L. (2008). *Cardinal Alignments and the Golden Section: Principles of Ancient Cosmography and Design*. Charleston, SC: Book-Surge.

Tuzzi, A., Popescu, I.-I., & Altmann, G. (2010a). The golden section in texts. *ETC – Empirical Text and Culture Research, 4*, 30–41.

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