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. $x, y \in <0, 1>$. The area between the bisector and the Lorenz curve is usually called Gini's coefficient, as can be seen in Figure 1.

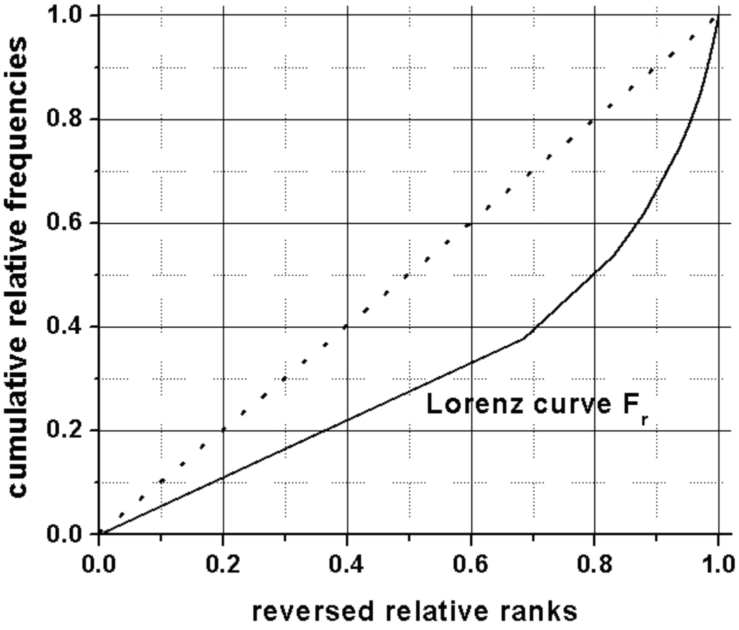

Figure 1. The Lorenz curve (from Popescu, I.-I. et al. 2009: 57)

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

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

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**	$\mathbf{R_4}$	**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

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

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

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

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.

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

For example, for the distribution in Table 1 we obtain

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

The *h*-point is defined as

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

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

$$(7) \qquad R = 1 - \frac{L_h}{L}$$

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

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

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

$$h = [3(3)\text{-}2(1)]/[3-2+3-1] = 7/3 = 2.3333.$$

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] $R_1 = 1-\left(F(h)-\dfrac{h^2}{2N}\right)$, 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:

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

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.