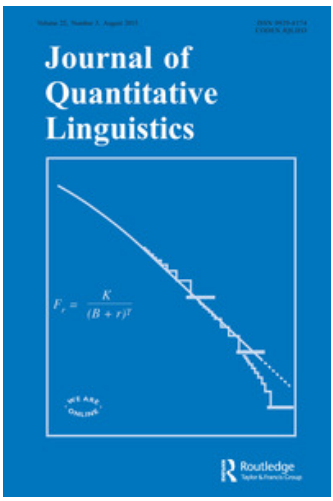

Journal of Quantitative Linguistics

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

Testing the Thematic Concentration of Text

Radek Čech[a], Radovan Garabík[b] & Gabriel Altmann[c]

[a] Department of Czech Language, University of Ostrava, Ostrava, Czech Republic

[b] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

[c] Lüdenscheid, Germany

Published online: 09 Jul 2015.

To cite this article: Radek Čech, Radovan Garabík & Gabriel Altmann (2015) Testing the Thematic Concentration of Text, Journal of Quantitative Linguistics, 22:3, 215-232, DOI: 10.1080/09296174.2015.1037157

To link to this article: http://dx.doi.org/10.1080/09296174.2015.1037157

Journal of Quantitative Linguistics, 2015
Vol. 22, No. 3, 215–232, http://dx.doi.org/10.1080/09296174.2015.1037157

Routledge
Taylor & Francis Group

Testing the Thematic Concentration of Text*

Radek Čech[a], Radovan Garabík[b] and Gabriel Altmann[c]
[a]Department of Czech Language, University of Ostrava, Ostrava, Czech Republic; [b]Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia; [c]Lüdenscheid, Germany

ABSTRACT

The aim of the article is to evaluate and address the limits of an existing approach to the analysis of the thematic concentration of text. To overcome these limits, the article proposes and applies both a modification of the measurement of thematic concentration – known as secondary thematic concentration and proportional thematic concentration – and methods for their statistical testing. The results show that the modification, as well as the application of the proposed tests, enhances the possibilities for analysing the thematic characteristics of text. The article uses 20 Slovak texts of the same genre written by one author.

INTRODUCTION

Every meaningful text, written or spoken, is produced with some goal or goals. Of course, there is an infinite number of these goals (for instance, the transmission of a message, a deliberate lie, a command, fun, "killing time", etc.) and an infinite number of ways to achieve them. Despite a huge variability of potential goals and means of their linguistic realization, texts (like human language as a whole) embody important regularities which can be viewed as a result of more general principles, such as the principle of least effort (Zipf, 1949) or self-regulation in a synergetic model of language (Köhler, 1986, 2005). These regularities can be captured, described, modelled mathematically and, in the best case, incorporated into a theory. Because there is, to our knowledge, no text theory in the sense of Bunge (1983), the majority of text analyses are either deliberately non-theoretical (specifically, computational linguists usually address practical problems and

*Address correspondence to: Radek Čech, Department of Czech Language, University of Ostrava, Reální 5, Ostrava 701 03, Czech Republic. E-mail: cechradek@gmail.com

they are not concerned with theoretical aspects) or they strive to reveal some characteristics of text and relationships among them in order to model some aspects of "text behaviour" (e.g. Wodak & Meyer, 2001; Wimmer, Altmann, Hřebíček, Ondrejovič, & Wimmerová, 2003; Hřebíček, 2007; Krippendorff, 2013). The present study should be viewed as a further step in the endeavour to explore text properties. Specifically, it is focused on methodological aspects of the analysis of the so-called thematic concentration of text. Thematic concentration (hereinafter *TC*) can be interpreted as a manifestation of the writer's or speaker's effort to communicate some topic(s) more intensively than other topics, or – importantly – more intensively than would be expected from 'neutral' language/text behaviour (cf. Section 2). Thus, the *TC* represents a regularity which appears despite a huge potential variability of means of 'manipulating' topic(s) of communication. Like any linguistic concepts, 'thematic concentration' is a definition-dependent concept.

Because of the huge variability of text characteristics, more complex methodological problems emerge in comparison to analyses focused on phonetics, morphology, lexicology or syntax. This fact leads us to explore thoroughly some aspects of analysis of the *TC*; first we focus on some limits of the existing approach, and then, as a consequence, we propose ways of overcoming these limits. Specifically, we present both modifications of the measurement of the *TC* and methods for their statistical testing. For the analysis we use 20 Slovak texts of the same genre written by one author (S. Svoráková) (see Appendix).

THEMATIC CONCENTRATION OF TEXT

The method of analysis of the *TC* was introduced by Popescu (2007) and elaborated by Popescu et al. (2009), Popescu and Altmann (2011) and Čech, Popescu and Altmann (2013). It was applied in textology by Sanada (2013), in literary theory by Wilson (2009), Davidová Glogarová, David, and Čech (2013), Davidová Glogarová and Čech (2013), in historical semantics by Čech (2013), and finally in an analysis of political speeches by Tuzzi, Popescu and Altmann (2010) and Čech (2014). By means of this method one can both identify words (or lemmas or co-referential units, such as hrebs) representing the main topic(s) of the text and quantify the author's concentration on the topic(s).

The method is based on two text characteristics: (1) the frequency distribution of words (or lemmas or co-referential units, such as hrebs) and

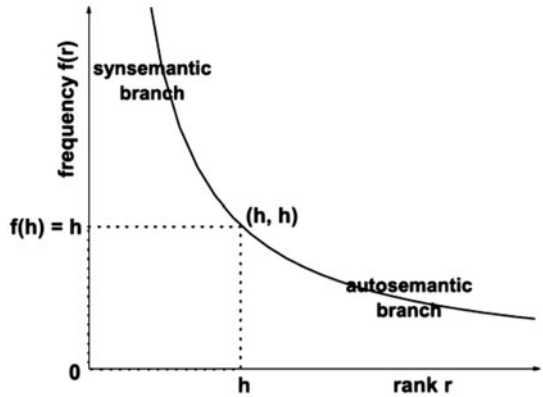

Fig. 1. A usual shape of the frequency distribution of words (or lemmas) in the majority of texts and an illustration of determination of the *h*-point (cf. Popescu et al., 2009, p. 17).

(2) the so called *h*-point (cf. Popescu, 2007). If one takes almost any text (exceptions are represented by Dadaistic texts, texts written by people with mental retardation or Wernicke's aphasia etc.) and ranks the words in order of decreasing frequency, one usually obtains a result such as that presented in Figure 1. The *h*-point, which is defined as a point where frequency equals rank (see Formula 1 below), separates in a fuzzy way the most productive synsemantics from autosemantics (see Figure 1, for more details, cf. Popescu et al., 2009, p. 17ff.). It is defined as

$$h = \begin{cases} r_i, & \text{if there is } r_i = f(r_i) \\ \frac{f(r_i)r_{i+1} - f(r_{i+1})r_i}{r_{i+1} - r_i + f(r_i) - f(r_{i+1})} & \text{if there is } r \neq f(r) \end{cases} \quad (1)$$

where r_i is a rank and $f(r_i)$ is the respective frequency of this rank; given that r_i is the highest number for which $r_i < f(r_i)$ and r_{i+1} is the lowest number for which $r_{i+1} > f(r_{i+j})$. Thus, if no rank is equal to the respective frequency, one computes the lower part of Formula (1) consisting of neighbouring values. Having stated the *h*-point, we consider all autosemantics occurring at lower ranks as thematic words because they signalize the frequent repetition of the given autosemantics.[1] In other words, the occurrence of autosemantics above the *h*-point (i.e. in the *synsemantic* branch) can be interpreted as some kind

[1]It should be mentioned that not all autosemantics need be considered to express the thematic properties of the text; for instance Popescu et al. (2009) use only nouns and their predicates of the first order, i.e. adjectives and verbs, for the analysis of the *TC*.

of anomaly in comparison to "neutral" texts which are not strongly concentrated on particular topic(s).

Let us symbolize the ranks and frequencies of these autosemantics as r' and $f(r')$ respectively. The thematic concentration is defined as

$$TC = 2\sum_{r'=1}^{T}\frac{(h-r')f(r')}{h(h-1)f(1)}, \tag{2}$$

where $f(1)$ is the highest frequency in the text and T is the number of autosemantics with $r < h$; if there are more words with the same frequency in the rank-frequency distribution, r' can also be represented by the average rank; for example, in Table 1, ranks 3, 4, and 5 may be re-ranked to 4 because the frequencies are equal, etc. Of course there also exist other possibilities for the quantification of the thematic characteristics of the text (cf. Čech, Garabík, & Altmann, forthcoming).

Additionally, it is a matter of fact that the study of word forms for this purpose is scarcely relevant, because the more analytical a language, the smaller is the number of forms. For instance, if a poet speaks only about his own feelings, in analytical languages "I" will appear many times, while in highly synthetic languages it may not appear at all as a separate word but only in the form of affixes. The problems of the relationship between the *TC* and language units are discussed by Popescu and Altmann (2011) and Čech, Popescu and Altmann (2013). The factor in the denominator of (2), $h(h-1)/2$, is the maximum given in the case that there are autosemantics at all ranks r'.

For illustration let us take the calculation of the *h*-point in the frequency distribution of words (in fact it is lemmas, i.e. canonical word forms, that are determined; for example the lemma *do* represents the word forms *do*,

Table 1. The eight most frequent lemmas in text No. 15. Thematic lemmas (i.e. autosemantics with $r \leq h$) are bolded.

Rank	Average rank	Lemma	Frequency
1	1	*v* [in]	19
2	2	*a* [and]	17
3	4	*byť* [be]	9
4	4	*jeho* [his]	9
5	4	***obraz* [picture]**	9
6	6	***rok* [year]**	8
7	7.5	*tento* [this]	6
8	7.5	*ako* [as]	6

does, *did*, *done*, and *doing*) in text No. 15; the rank-frequency distribution of the eight most frequent lemmas is presented in Table 1.

Since in Table 1, r ≠ *f*(*r*), for the computation of the *h*-point we use the lower part of Formula (1), i.e.

$$h_{text15} = \frac{8(7) - 6(6)}{7 - 6 + 8 - 6} = 6.6667.$$

There are two autosemantics with $r < h$ in Table 1 (*obraz* [picture], *rok* [year]). Thus, the *TC* of this text is computed as follows (average rank is used for the computation):

$$TC_{text\,15} = 2\left(\frac{(6.6667 - 4)9}{6.6667(6.6667 - 1)19} + \frac{(6.6667 - 6)8}{6.6667(6.6667 - 1)19}\right) = 0.081734.$$

In order to compare the given texts with one another or with other text (-type)s, one can use the theoretical variance of *TC*, defined as (cf. Popescu & Altmann, 2011):

Table 2. Thematic concentrations (*TC*), their variances Var(*TC*) and the lengths (*N*) of the analysed texts.

Text	***h***	***TC***	**Var(*TC*)**	***N***
1	8.33	0.002584	0	750
2	11.5	0.056522	0	1084
3	11	0.025253	0.00000608	998
4	8.5	0.030166	0	631
5	9	0.028935	0	618
6	8	0.144599	0.00000649	765
7	7.5	0.100513	0	594
8	9.5	0.013313	0.00000200	1094
9	9	0.049383	0	807
10	9	0.005435	0	702
11	6.33	0	0	448
12	6.5	0	0	403
13	9.5	0.074303	0.00001840	748
14	5.5	0	0	249
15	6.67	0.081734	0.00013153	402
16	5	0	0	228
17	7	0	0	397
18	7	0.059524	0	460
19	13	0.130738	0.00000510	2075
20	12	0.018218	0.00000204	1218

$$Var(TC) = \left(\frac{2}{h(h-1)f(1)}\right)^2 \left(\sum_{r'=1}^{T} f(r')\right) m_{2r}, \tag{3}$$

where $m_{2r'}$ is the variance (the second central moment) of thematic words above the h-point, i.e.

$$m_{2r} = \frac{\sum_{r'=1}^{T} (r' - m_{1r})^2 f(r')}{\sum_{r'=1}^{T} f(r')}, \tag{4}$$

where $m_{1r'}$ is the first central moment, i.e.

$$m_{1-r} = \frac{\sum r' \cdot f(r')}{\sum f(r')} \tag{5}$$

All *TC*-values and their variances for 20 texts by Svoráková are presented in Table 2.

METHODS FOR MEASURING AND STATISTICAL TESTING OF THE *TC*

For comparing individual texts the use of the asymptotic u-test was proposed by Popescu and Altmann (2011); it is defined as

$$u = \frac{|TC_1 - TC_2|}{\sqrt{Var(TC_1) + Var(TC_2)}}. \tag{6}$$

However, if we try to use Formula (6) for the data in Table 1, some problems may emerge:

(1) there are texts with $TC = 0$; (2) there is frequently only one thematic word in the pre-h-domain (which means that $Var(TC)$ equals zero).

As regards the first problem, the $TC = 0$ can be easily interpreted as a manifestation of the thematic "neutrality" of the text. However, this does not seem to be a very practical solution if one wants to analyse thematic differences among texts. To solve this disadvantage, it is possible to start with the h-point and its theoretical interpretation; it is stated that the h-point represents a *fuzzy* border between synsemantic and autosemantic words (see Figure 1). Consequently, from a theoretical point of view there is no problem with doubling the h-point; specifically, this means that h is multiplied by two in Formula (2), and we obtain the so-called secondary thematic concentration (*STC*)

$$STC = \sum_{r'=1}^{2h} \frac{(2h - r')f(r')}{h(2h-1)f(1)}. \quad (7)$$

Consequently, it is necessary to modify the variance

$$Var(STC) = \frac{\left[\sum_{r'=1}^{T} f(r')\right] m_{2r'}}{[h(2h-1)f(1)]^2}, \quad (8)$$

where $m_{2r'}$ is the variance of the autosemantics with $r < 2h$ (see Formula (4)). This approach is mentioned only marginally as a possibility in Popescu et al. (2009, p. 103); however, to our knowledge it has not yet been used in any analysis. The obvious advantage of this approach is that the probability that some autosemantics appear above $2h$ is much higher. The results of the *STC* for 20 texts by Svoráková are presented in Table 3.

If one observes Table 3, one can see that the adoption of the *STC* eliminates problem (1) totally (all $STC > 0$) and problem (2) in 17 instances (texts No. 11, 16 and 17 have Var(*STC*) = 0, because there are thematic

Table 3. The secondary thematic concentrations (*STC*) and their variances Var(*STC*) of the analysed texts.

Text	**2*h***	***STC***	**Var(*STC*)**
1	16.67	0.016529	0.00000936
2	23	0.061166	0.00001828
3	22	0.057299	0.00001478
4	17	0.059389	0.00005074
5	18	0.086329	0.00003596
6	16	0.118699	0.00002701
7	15	0.083333	0.00005357
8	19	0.047515	0.00000326
9	18	0.090414	0.00008502
10	18	0.069764	0.00001924
11	12.67	0.030538	0
12	13	0.061086	0.00002194
13	19	0.118177	0.00007526
14	11	0.065035	0.00004315
15	13.34	0.091323	0.00000694
16	10	0.051282	0
17	14	0.028846	0
18	14	0.102647	0.00008056
19	26	0.101994	0.00001516
20	24	0.065649	0.00000400

words with the same average rank in their frequency distribution). This means that it is possible to test the differences of *STC* among all texts except the differences among texts No. 11, 16 and 17. There is no doubt that this is an important benefit in comparison to Table 2.

As for the second problem – the occurrence of only one thematic word in the pre-*h*-domain, which means that *Var*(*TC*) equals zero and consequently it is not possible to test differences by means of formula (6)[2], in such cases one either can apply the *STC* (see above) or use a different approach. To follow the second strategy, we propose proportional thematic concentration (*PTC*) and two tests for comparing the *PTC* in two texts.

Let the proportion of thematic words in the pre-*h*-domain be *PTC*, computed as

$$PTC = \frac{1}{N_h} \sum_{r' < h} f(r'), \tag{9}$$

where N_h = frequency of all words r_1, ... , r_h, i.e. all words in the pre-*h*-domain, and the sum of $f(r')$ is the frequency of all autosemantic words occurring in the pre-*h*-domain; the variance of *PTC* is

$$Var(PTC) = \frac{PTC(1 - PTC)}{N_h}. \tag{10}$$

The asymptotic normal test now yields

$$u = \frac{|PTC_1 - PTC_2|}{\sqrt{Var(PTC_1) + Var(PTC_2)}}. \tag{11}$$

As an example, consider text No. 18, in which $N_h = 77$ and in which there is only one autosemantic occurring in the pre-*h*-domain with $f(r') = 11$, hence

$$PTC_{text18} = \frac{11}{77} = 0.1429$$

and

$$Var(PTC_{text18}) = \frac{0.1429(1 - 0.1429)}{77} = 0.001591.$$

[2]Of course, theoretically, it is possible to test differences of the *TC* between two texts, if one of the text has Var(*TC*) > 0 and the other Var(*TC*) = 0.

We compute analogically for text No. 7, in which $N_h = 97$ and in which there is only one autosemantic occurring in the pre-h-domain with $f(r') = 14$; we obtain $PTC_{text7} = 0.1443$ and $Var(PTC_{text7}) = 0.001273$. Now we can compare the thematic concentrations of these texts by means of Formula (7)

$$u = \frac{|0.1429 - 0.1443|}{\sqrt{0.001591 + 0.001273}} = 0.02,$$

which means non-significant difference (for the significance level $\alpha = 0.05$, $u \geq 1.96$).

If we want to perform an exact test, we consider the smaller of the two *PTC*-s as the theoretical value, and using the data from the other text, i.e. N_{h2} and $x = \sum_{r'_2 < h} f(r'_2)$ we compute

$$P(X \geq x) = \sum_{j \geq x} \binom{N_{h2}}{j} p'^{j} q'^{N-j},$$

Table 4. The proportional thematic concentrations (*PTC*) and their variances Var(*PTC*) of the analysed texts.

Text	***h***	***PTC***	**Var(*PTC*)**
1	8.33	0.066176	0.00045439
2	11.5	0.088983	0.00034350
3	11	0.134078	0.00064861
4	8.5	0.086207	0.00067910
5	9	0.093458	0.00079181
6	8	0.241611	0.00122977
7	7.5	0.144330	0.00127318
8	9.5	0.103960	0.00046115
9	9	0.100000	0.00075000
10	9	0.084112	0.00071997
11	6.33	0	0
12	6.5	0	0
13	9.5	0.185629	0.00090521
14	5.5	0	0
15	6.67	0.239437	0.00256488
16	5	0	0
17	7	0	0
18	7	0.142857	0.00159025
19	13	0.192623	0.00031869
20	12	0.107280	0.00036694

where $p' = PTC$ and $q' = (1 - PTC)$. If this probability is smaller than, say, 0.05, we consider the difference as significant. For example, in text No. 18 we had $p' = 0.1429$; in text No. 7 we had $N_h = 97$ and $x = 14$. Hence computing (9) we obtain

$$P(X \geq 14) = \sum_{j \geq 14} \binom{97}{j} 0.1429^j (1 - 0.1429)^{97-j} = 0.5279,$$

telling us that there is no difference between the two texts in the sense of the *TC* (seen from this point of view). The results of the *PTC* for 20 texts by Svoráková are presented in Table 4.

Obviously, these tests enhance the possibilities for the analysis of the *TC* (cf. Table 2). However, some texts are still no statistically comparable because of zero values of the *PTC*. Therefore, we apply the *PTC* not only for autosemantics in pre-*h* domain, but also in pre-2 *h* domain. Let us call this index secondary proportional thematic concentration (*SPTC*). The results of the *SPTC* for 20 texts by Svoráková are presented in Table 5.

Table 5. The secondary proportional thematic concentrations (*SPTC*) and their variances Var (*SPTC*) of the analysed texts.

Text	**2 *h***	***SPTC***	**Var(*SPTC*)**
1	16.67	0.078534	0.00037888
2	23	0.166163	0.00041859
3	22	0.202952	0.00059691
4	17	0.263804	0.00119148
5	18	0.267857	0.00116732
6	16	0.284314	0.00099745
7	15	0.142857	0.00087464
8	19	0.136029	0.00043208
9	18	0.379487	0.00120757
10	18	0.246988	0.00112039
11	12.67	0.096386	0.00104934
12	13	0.238636	0.00206465
13	19	0.272340	0.00084328
14	11	0.261538	0.00297132
15	13.34	0.157407	0.00122806
16	10	0.218182	0.00310143
17	14	0.151515	0.00129857
18	14	0.300813	0.00170996
19	26	0.288026	0.00033182
20	24	0.180593	0.00039887

The *SPTC* is the method which allows to compare statistically all texts in the sample.

DIFFERENCES DUE TO DIFFERENT MEASUREMENTS OF THE THEMATIC CONCENTRATION

It is well known in statistics that different statistical tests can yield different results. The same is true for the use of different methods of measurement – in our case the *TC*, *STC*, *PTC*, and *SPTC*. Consequently, for an appropriate interpretation of particular methods both a comparison of the methods and an observation of differences of results (if they occur) are necessary. In other words, the methods presented in this article should be focused on the same text property; to interpret this property, one has to know the aspects of the applied methods as well as possible.

For a comparison of the methods the correlation coefficient was used, see Table 6 and Figures 2, 3, 4, 5, 6, 7.

Table 6. Correlation coefficients (r) between particular indicators.

	r	R^2
TC – PTC	**0.8584**	0.7369
TC – STC	**0.7897**	0.6236
STC – PTC	**0.7583**	0.5750
SPTC – STC	**0.7044**	0.4962
SPTC –TC	0.3308	0.1094
SPTC –PTC	0.2450	0.0600

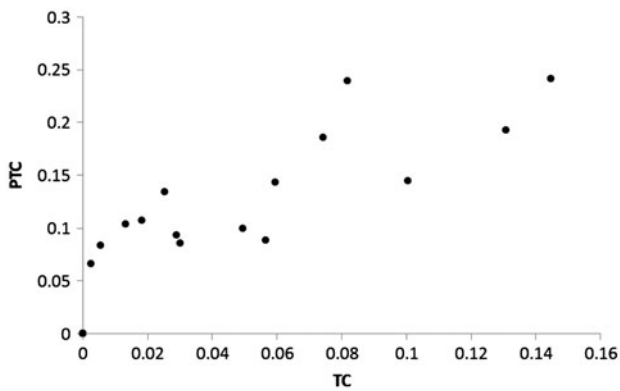

Fig. 2. Relationship between the *TC* and *PTC* in the analysed texts.

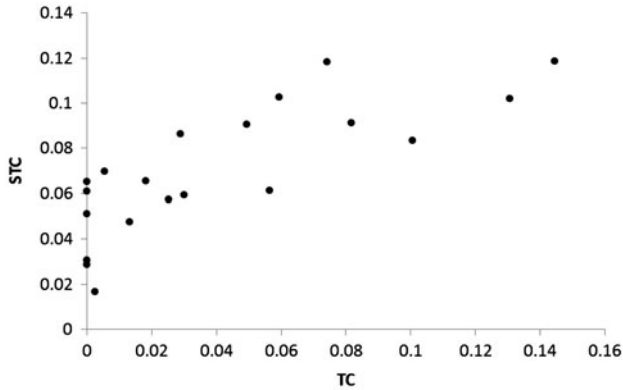

Fig. 3. Relationship between the *TC* and *STC* in the analysed texts.

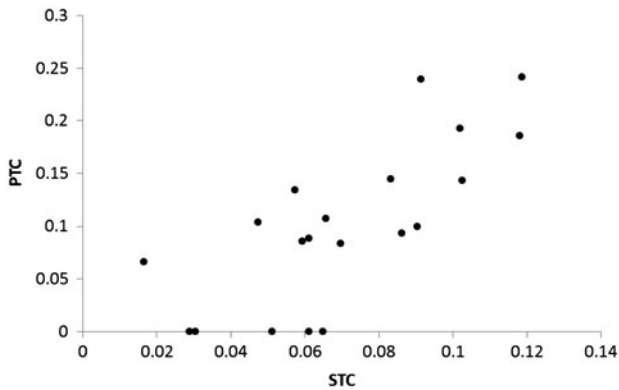

Fig. 4. Relationship between the *STC* and *PTC* in the analysed texts.

Statistically significant values are bolded; at significant level 0.05. R^2 expresses the determination coefficient. Pairs of indicators are ranked in decreasing order in accordance to the coefficient of determination.

The results reveal significant correlation between indicators as follows: the *TC* and *PTC*, *TC* and *STC*, *STC* and *PTC*, *SPTC* and *STC*; non-significant correlation between both pairs the *STPC* and *TC* and between *STPC* and *PTC*. Further, even though the correlation between the *SPTC* and *STC* is significant, the low coefficient of determination indicates weaker correlation

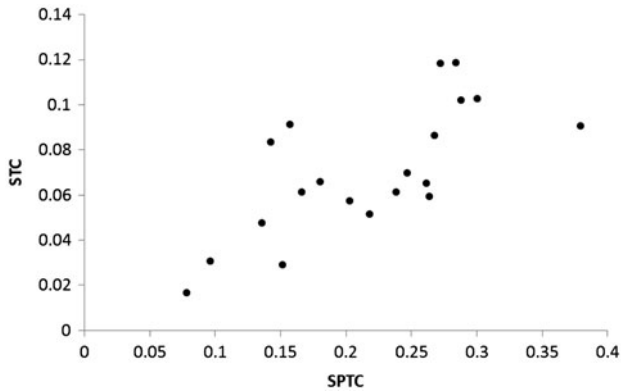

Fig. 5. Relationship between the *SPTC* and *STC* in the analysed texts.

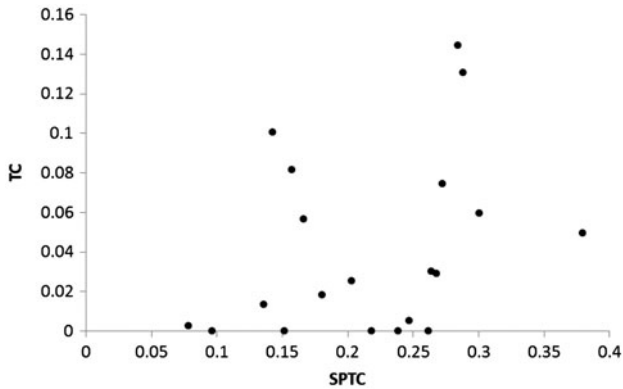

Fig. 6. Relationship between the *SPTC* and *TC* in the analysed texts.

with regard to the other significant correlations. Consequently, the *STCP* does not seem to capture the same property as the other methods and is not proper for the analysis of thematic characteristics of text.

A closer observation of the results reveals a specific tendency for the relationship between the *TC* and *STC*. Particularly, for texts with the highest *TC*, $STC < TC$, while for texts with the lower *TC*, $STC > TC$ (cf. Figure 8).

This finding is not surprising, if one realizes the properties of these particular measurements. Specifically, even though the *STC* captures more

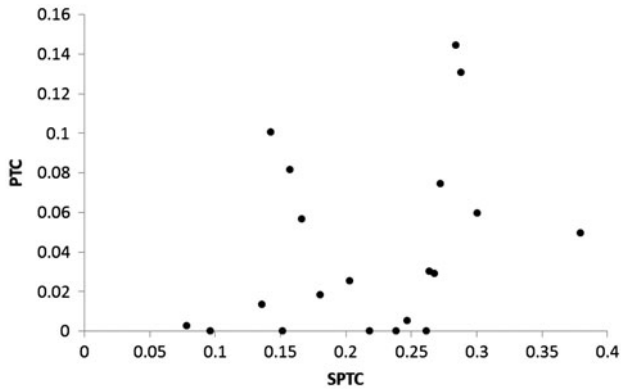

Fig. 7. Relationship between the *SPTC* and *PTC* in the analysed texts.

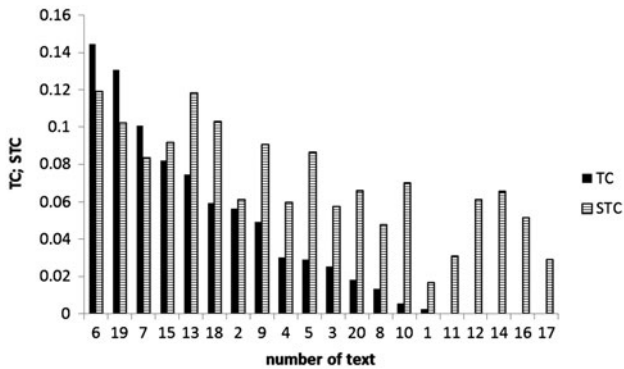

Fig. 8. *TC* and *STC* in particular texts. Texts are ranked (*x*-axis) in decreasing order in accordance to *TC*.

thematic words and consequently there should be a tendency $STC > TC$, the normalization, $h(2h-1)f(1)$, can also cause the opposite, i.e. $STC < TC$. A closer observation of texts with $STC < TC$ shows that the high *TC* is caused by word(s) with extremely low rank and high frequency (with regard to the *h*-point); for instance, text No. 6 ($h = 8$, $f(1) = 41$) contains thematic words with $r = 3$, $f(r) = 22$ and $r = 4$, $f(r) = 14$. Consequently, a comparison of *TC* and *STC* can be used as an indicator of the extremeness of the *TC*; therefore, texts with $STC < TC$ can be considered as extremely concentrated texts.

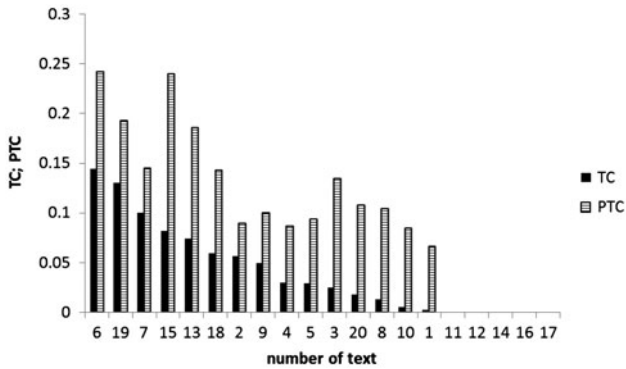

Fig. 9. *TC* and *PTC* in particular texts. Texts are ranked (*x*-axis) in decreasing order in accordance to *TC*.

As for the other relationships between indicators, no similar tendency emerged; as an example, the relationship between the *TC* and *PTC* is presented in Figure 9.

5. CONCLUSION

This article has presented four possible ways of measuring thematic concentration and proposed various tests for comparing texts. Having analysed one author, the result cannot be considered general. Nevertheless, it represents a possible starting point for further investigations. Short texts have a strong proneness to variation, but it may also be the conscious intention of the author to concentrate the content of the text. In order to enhance the power of this research, we plan to propose various other definitions of concentration based on frequencies, sequences and hrebs (cf. Čech, Garabík, & Altmann, forthcoming).

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

FUNDING

This work was supported by the Czech Science Foundation [grant number P406/11/0268].

REFERENCES

Bunge, M. A. (1983). *Treatise on Basic Philosophy: Epistemology & Methodology*. Dordrecht: Springer.

Čech, R. (2013). Why should be quantitative in historical semantics and textology?. In J. David, R. Čech, L. Radková, J. Davidová Glogarová, & H. Šústková (Eds), *Slovo a text v historickém kontextu - perspektivy historickosémantické analýzy jazyka* [Word and text in a historical context. Perspectives of historical and semantic analysis of language] (pp. 32–34). Brno: Host. (in Czech)

Čech, R. (2014). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949-2011). *Quality & Quantity*, *48*(2), 899–910.

Čech, R., Garabík, R., & Altmann, G. (forthcoming). Some new indicators of thematic concentration.

Čech, R., Popescu, I. I., & Altmann, G. (2013). Methods of analysis of a thematic concentration of the text. *Czech and Slovak Linguistic Review*, *3*, 4–21.

Davidová Glogarová, J., & Čech, R. (2013). Tematická koncentrace textu – některé aspekty autorského stylu Ladislava Jehličky [Thematic concentration of a text – some aspects of Ladislav Jehlička's authorial style]. *Naše řeč*, *96*, 234–245. (in Czech).

Davidová Glogarová, J., David, J., & Čech, R. (2013). Analýza tematické koncentrace textu – komparace publicistiky Ladislava Jehličky a Karla Čapka [Analysis of the thematic concentration of texts: A comparison of journalistic texts by Ladislav Jehlička and Karel Čapek]. *Slovo a slovesnost*, *74*, 41–54. (in Czech).

Hřebíček, L. (2007). *Text in Semantics*. Prague: Oriental Institute.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik* [On synergetic linguistics. Structure and dynamics of lexicon]. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds), *Quantitative Linguistics. An International Handbook* (pp. 760–774). Berlin, New York: de Gruyter.

Krippendorff, K. (2013). *Content Analysis. An Introduction to Its Methodology*, (3rd edition). Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE Publications, Inc.

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* (pp. 555–566). Berlin, New York: Mouton de Gruyter.

Popescu, I.-I., & Altmann, G. (2011). Thematic concentration in texts. In E. Kelih, V. Levickij, & Y. Matskulyak (Eds), *Issues in Quantitative Lingustics*, Vol. 2 (pp. 110–116). Lüdenscheid: RAM-Verlag.

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

Sanada, H. (2013). Thematic concentration in Japanese prose. In I. Obradovic, E. Kelih, & R. Köhler (Eds), *Methods and Applications of Quantitative Linguistics. Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)* (pp. 130–140). Belgrade, Serbia, April 26–29, 2012. Belgrade: University of Belgrade.

Tuzzi, A., Popescu, I.-I., & Altmann, G. (2010). *Quantitative Analysis of Italian Texts* Lüdenscheid: RAM-Verlag.
Wilson, A. (2009). Vocabulary richness and thematic concentration in internet fetish fantasies and literary short stories. *Glottotheory*, *2*(2), 97–107.
Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., & Wimmerová, S. (2003). *Úvod do analýzy textov* [Introduction to text analyses]. Bratislava: Veda.
Wodak, R., & Meyer, M. (Eds). (2001). *Methods of Critical Discourse Analysis* Los Angeles: Sage Publications.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.

TEXTS BY S. SVORÁKOVÁ

Text 1. **Svoráková, S**. Čakanie na Štraussa Review of: Tomáš, Štrauss, Metamorfózy umenia XX. storočia. Bratislava: Kalligram, 2001. In: *Dart - Revue súčasného výtvarného umenia*. 10/2003. p. 37.
Text 2. **Svoráková, S.** Dvojhlasné dejiny a univerzálna kultúra? Review of: Mária, Orišková: Dvojhlasné dejiny umenia. Bratislava: Petrus, 2002. In: *Dart - Noviny o súčasnom výtvarnom umení* 02 /2003. p. 3.
Text 3. **Svoráková, S.** Stratená moderna Review of: Tomáš, Štrauss: Zo seba vystupujúce umenia. Príspevok k stratifikácii stredoeurópskych avantgárd. Bratislava: Kalligram, 2003. In: D*art - Noviny o súčasnom výtvarnom umení* 01/2004. p. 3.
Text 4. **Svoráková, S.** Znovuobjavené klenoty Review of: Ján Hollý – Emil Makovický: Selanky. (Úvodný text K. Szmudová). Banská Bystrica: Štátna vedecká knižnica, 2007. In: *Slovenské pohľady*. 7- 8/ 2007, p. 276 -277.
Text 5. **Svoráková, S.** Smrť jej nepristane Review of: Nová krv. (Úvodný text I. Jančár). Bratislava: Galéria Mesta Bratislavy, 2008. In: *Literárny (dvoj)týždenník*. 5- 6/ 2009, p. 13.
Text 6. **Svoráková, S.** Ruská interpretácia slovenského naturizmu Review of: Alla, Mašková: Slovenský naturizmus v časopriestore. (Prel. Hedviga Kubišová) Bratislava, 2009. In: *Literárny (dvoj)týždenník*. 13-14/2011, p. 12.
Text 7. **Svoráková, S.** ...a poslední nie sú prví – Na margo výstavy Review of the exposition: Bienále v čase normalizácie v Stredoslovenskej galérii v B. Bystrici. In: *Literárny (dvoj)týždenník*, 37-38/ 2011, p. 13.
Text 8. **Svoráková, S.** Voľným okom – List zo Slovenska Review of: Voľným okom. (Úvodný text Ľ. Hološka). Martin: Vydavateľstvo Matice slovenskej, s. r. o., 2006. In: *Mecenat i mir - Literarno-chudožestvennyj I kulturnyj magazin.* No 41- 42- 43- 44. Moskva, 2009. p. 356- 358.
Text 9. **Svoráková, S.** Alternatívy slovenskej grafiky. Review of an exposition. In: *Literárny týždenník*. 5/1998, p. 14.
Text 10. **Svoráková, S.** Veľké ambície malej grafiky. Review of an exposition: XIV. Ročník Medzinárodného trienále drevorezu a drevorytu. In: *Literárny týždenník*. 6/1999, p.14.
Text 11. **Svoráková, S.** 200 plechoviek Campbellovej polievky. In: *Literárny (dvoj) týždenník)* 26-27/ 2004, p. 14.
Text 12. **Svoráková, S.** Plenér Liptov 1999. In: Plenér Liptov, (1999). Úvodný text v katalógu Medzinárodného sympózia (p. 1999). Banská Bystrica: Akadémia umení.

Text 13. **Svoráková, S.** Plenér Liptov 2001. In: *Plenér Liptov 2001*. Úvodný text v katalógu Medzinárodného sympózia Vyd.: Norami pre Galériu P&P, Mesto Liptovský Mikuláš, Rotary Club Liptovský Mikuláš a FVU Bratislava, 2001.
Text 14. **Svoráková, S.** Kabaret života. In: *Kabaret života – Kamila Štanclová: Obrazy, grafika.* Zvolen: Vlastivedné múzeum, 1991.
Text 15. **Svoráková, S.** Štefan Prukner Bartušek: Mágia obrazu. In: *Originál.* 3/1998, p. 8.
Text 16. **Svoráková, S.** Margita. In: *Margita*. Úvodný text katalógu Jaroslava Uhela. 2000.
Text 17. **Svoráková, S.** Štefan Prukner Bartůšek v zahraničí a doma. In: *Slovenská republika*. 8.9.1998, p. 10.
Text 18. **Svoráková, S.** Národná múza Ladislava Dunajského. In: *Priekopník.* 10.4.1997, p. 3.
Text 19. **Svoráková, S.** Majstrovstvo bulharských ikon. In: *Výtvarný život*. 3/1990, p. 60-65.
Text 20. **Svoráková, S.** Nahlas o jednom areáli – Pamätník SNP po novej úprave. In: *Priekopník.* 10.4.1990, p. 3.