Mathematical Linguistics
2016, Volume 2, Number 1, pp. 5-13

Text length and the thematic concentration of text

Radek Čech[1]
University of Ostrava
cechradek@gmail.com

Miroslav Kubát
University of Ostrava
miroslav.kubat@gmail.com

ABSTRACT

The impact of text length very often biases results of stylometric indices which are based on rank-frequency distribution (e.g. type-token ratio, repeat rate, entropy). The aim of the article is to observe the relation between text size and thematic concentration indicators (*TC*, *STC*). The corpus consists of 1471 English texts of various genres. The obtained results show that thematic concentration is independent of text length in the interval <200; 6500>. Given that the analysis corroborates the findings of the previous research in Czech language, *TC* and *STC* seem to be reliable stylometric indicators applicable to text analyses of different languages.

Keywords: Normalization, Stylometry, Text analysis, Text length, Thematic concentration

[1]- Corresponding author: Department of Czech language, University of Ostrava, Czech Republic.

INTRODUCTION

In quantitative text linguistics, the text length is a factor which influences a majority of methods used for text analysis. The most illustrative is a history of measurement of vocabulary richness; this history can be interpreted as a series of attempts to eliminate an undesirable impact of text length. Typically, some normalization in a computation is used (cf. Popescu et al. 2009, 2011; Wimmer et al. 2003; Těšitelová 1992; Guiraud 1954). The main problems of normalization are as follows; a) particular ways of normalization are not linguistically interpretable (for instance, a use of logarithm or square root in a formula), b) in most cases the normalization does not work well – it somewhat decreases the impact of the text length on a given method, however, it still influences the measurement (cf. Čech 2015). Another way is to use only a part of the text (e.g. first 100 or 1000 words). But this approach does not respect the "integrity" of text and, consequently, it is not adequate, at least from some points of view. Recently, a method called the moving average type-token ratio (introduced by Convington and McFall 2010 and further elaborated by Kubát and Milička 2013) seems to bring promising results in the elimination of text length in the measurement of vocabulary richness. Even though this method is theoretically applicable not only to the measurement of type-token ratio, it has not been employed to other methods yet.

Despite the fact that the text length is indeed a very strong factor influencing particular indices (e.g., the type-token ratio, repeat rate, entropy, lambda structure of text), there are some methods which could be immune to it because of their character. For instance, there is no reason to expect that text activity (cf. Zörnig 2015) should decrease/increase with increasing text length. Further, in some cases, it is possible to determine empirically an interval in which indices are independent of the text length and, importantly, to give theoretical reasons for the determination (cf. Čech 2015).

The later approach seems to be acceptable for the measurement of thematic concentration, as was presented for Czech by Čech (2016). Needless to say, it is necessary to analyse more languages and more texts to explore the

relationship between the thematic concentration and text length. Therefore, we analyse English texts in this study. The article is organized as follows; two methods of measurement of thematic concentration are presented in Section 2, language material used for the analysis is described in Section 3, results are presented in Section 4 and, the article is finalized by Conclusion (Section 4).

METHODS OF MEASUREMENT OF THEMATIC CONCENTRATION

There were introduced several methods for the measurement of thematic concentration in linguistics (Popescu et al. 2009; Čech et al. 2013, 2015). For the purpose of this study, we use the original measurement of thematic concentration (hereinafter *TC*) (Popescu et al. 2009) and its modification called the secondary thematic concentration (hereinafter *STC*) (Čech et al. 2013). They are defined[1] as

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

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

where $f(1)$ is the highest frequency in the text, h is the h-point (see formula 3) 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). h-point is defined as

$$(3) \quad h = \begin{cases} r_i, & \text{if there is } r_i = f(r_i) \\ \dfrac{f(r_i)r_{i+1} - f(r_{i+1})r_i}{r_{i+1} - r_i + f(r_i) - f(r_{i+1})} & \text{if there is } r \neq f(r) \end{cases} \ ,$$

[1] Here, we present only formulas; for theoretical aspects of the approach, see references.

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 (3) consisting of neighboring values.

LANGUAGE MATERIAL

For the analysis, 1471 English texts of various genres were used, see Table 1. The length of texts lies in an interval N = <51; 360 000> tokens. All texts were processed by the software *QUITA* (Kubát et al. 2014). For a computation, word forms were used as a unit.

Table 1. Number of texts of different genres.

genre	**number of texts**
news	150
US presidential speeches	57
scientific texts	60
short stories	103
chapters of novels	777
novels	23
poems	124
personal letters	177

RESULTS

Before presenting the results, let us consider theoretical aspects of the relationship between the indices (*TC*, *STC*) and the text length. As for extremely short texts (e.g. poems, letters, tweets), the author has a great possibility to control the frequency characteristics of used vocabulary. However, minimal changes of frequency (in an extreme cases only one occurrence or its absence) influence the value of the indices fundamentally. For instance, a short letter (it contains N = 45 tokens) written by John Keats has the maximal value of the TC = 1; the absence of sole occurrence of word “Dilke” would mean that the *TC* of the

letter equals zero, see Table 2. Consequently, it does not seem to be adequate to use the method for the analysis of very short texts. A minimum length of text can be derived empirically, as is presented below.

Table 2. The rank-frequency distribution of word forms in a short letter (N = 45 tokens) written by John Keats.

Rank	**token**	**frequency**
1	*Dilke*	3
2	*of*	2
3	*this*	2
4	*shall*	2
5	*my*	2
6	*given*	1
…	…	…
45	*send*	1

The increasing length of text leads to the increasing frequency of synsemantics (caused by the grammar) and to the increasing value of the normalizing constant (the divisor in the formula 1 and 2) which would cause the decrease of the *TC* (or *STC*). On the other hand, the increasing length of text leads to increasing value of the *h*-point, consequently, the probability of occurrence of autosemantics above the *h*-point increases, too, and, consequently, it would cause the increase of the *TC* (or *STC*). Thus, there seem to be two opposite mechanisms whose interaction results in the elimination of the impact of text length on the thematic concentration.

As for extremely long texts (e.g, novels consisting of several volumes), there is a minimal chance that the author can control the frequency characteristics of words. The proportion of words is probably a result of some background mechanism, cf. "The longer the text, the more the writer loses his subconscious control over some proportions and keeps only the conscious control over contents, grammar, his aim, etc. But as soon as parts of 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“ (Popescu et al. 2012, 126–127).

The results (see Figure 2–4) confirm the theoretical assumptions mentioned above: the shortest texts have either zero values of indices, or extreme ones. One the other hand, the *TC* and *STC* of very long texts lies in a minimal interval.

For Czech, Čech (2016) derived empirically the interval in which the *TC* and *STC* are independent on the text length; specifically, the interval is N = <200; 6500> tokens. Analogously, we applied the same procedure and the results are presented in Figures 3 and 4. The linear regression line is almost horizontal, consequently, we can state that the *TC* and *STC* are independent on text length in the interval. Obviously, the boundaries of the interval shall be taken as fuzzy ones.

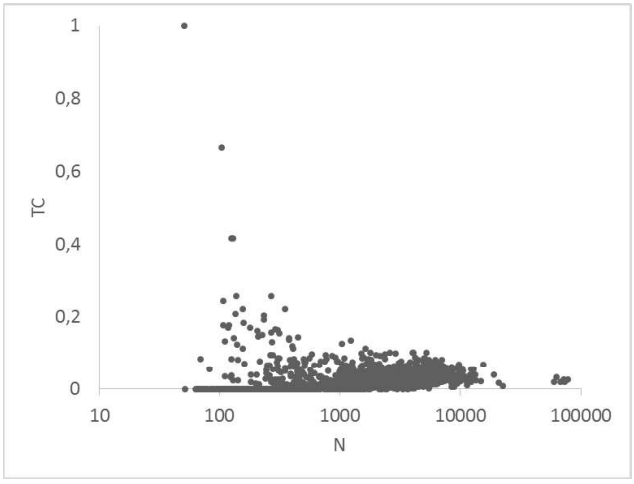

Figure 1. The relationship between the *TC* and text length in 1471 English texts. The *x*-axis is logarithmic.

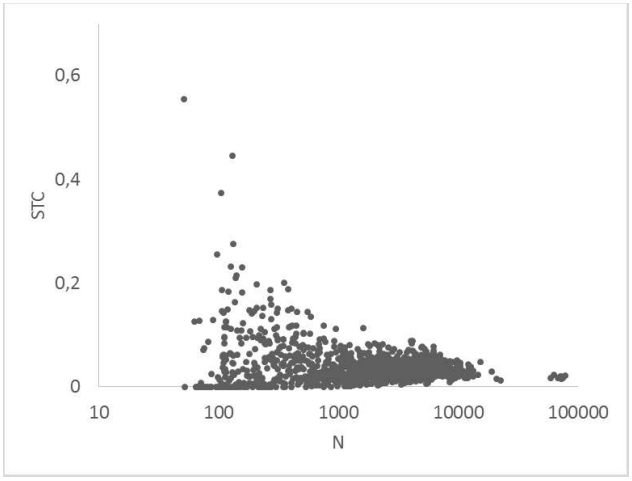

Figure 2. The relationship between the *STC* and text length in 1471 English texts. The *x*-axis is logarithmic.

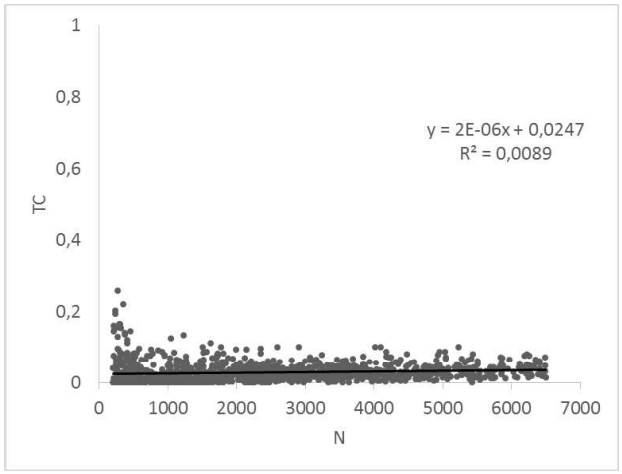

Figure 3. The relationship between the *TC* and text length in 1471 English texts with the length N = <200; 6500> tokens.

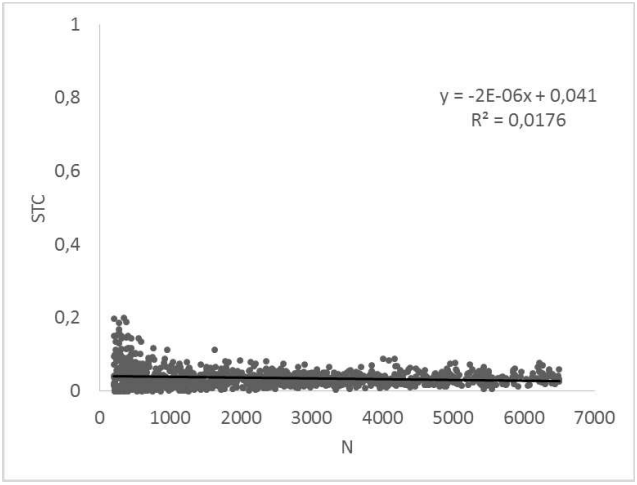

Figure 4. The relationship between the *STC* and text length in 1471 English texts with the length N = <200; 6500> tokens.

CONCLUSION

The study revealed that the values of both the *TC* and *STC* are independent on text length in the interval N = <200; 6500>. Thus, the indices can be "safely" used in this interval for any text analysis of English and Czech texts (cf. Čech 2016 for Czech). We assume that the same results could be obtained for other languages as well because there is no reason to expect some language specific characteristics, i.e., different boundary conditions. Needless to say, this prediction must be corroborated (or rejected) empirically.

REFERENCES

Čech, R. (2015). Text length and the lambda frequency structure of the text. In Mikros, G. K., Mačutek, J. (eds.) Sequences in language and text. De Gruyter, 71–88.

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.

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

Kubát, M., Milička, J. (2013). Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics 20(4), 339–349.*

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. (2011). The lambda-structure of texts. Lüdenscheid: RAM.

Těšitelová, M. (1992). Quantitative linguistics. Praha: Academia.

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

Zörnig, P. et al. (2015). Descriptiveness, Activity and Nominality in Formalized Text Sequences. Lüdenscheid: RAM.