Thematic Concentration and Vocabulary Richness

*Kubát Miroslav, Čech Radek*

University of Ostrava, Czech Republic

Abstract. The contribution investigates a relation between two stylometric features with promising results in text classification: thematic concentration and vocabulary richness. Namely secondary thematic concentration (*STC*), moving average type-token ratio (*MATTR*), and repeat rate (*RRMC*) are analysed. The main aim is to test the hypothesis that vocabulary richness negatively correlates with thematic concentration. The research is based on a corpus of more than 900 English texts from various genres. This study follows up a similar analysis (Čech 2016) which investigated Czech texts.

*Keywords: thematic concentration, vocabulary richness, correlation, type-token ratio, repeat rate*

Introduction

Several stylometric indices such as thematic concentration (Popescu et al. 2009), lambda-structure of text (Popescu et al. 2011), moving average type-token ratio (Covington, McFall 2010), nominality of text (Zörnig et al. 2016), or writer’s view (Popescu, Altmann 2007) have been proposed in recent years. It seems reasonable to assume systematic relationships among these indices because they express text characteristics which are an output of a predictable (by means of a statistical hypothesis) verbal behaviour. Specifically, if majority of these indices are useful tools for a text classification, i.e. they are able to detect systematic properties of language production, they should be governed by the similar prin­ciples or mechanisms. It is a great challenge for the text linguistics to reveal these principles and, finally, to develop a text theory which could explain human language behaviour with regard to the text characteristics. Because there is no text theory of this kind, we can try to extend our knowledge of general text properties by an analysis of relationships among particular indices. This approach leads not only to better understanding of the indices but also it can be an important step in the theory building.

 In this paper, we analyse the relationship between thematic concentration and vocabulary richness. These text properties have been analysed in several studies with promising results in terms of stylometry (e.g. Kubát, Čech 2016; Čech 2014; Popescu et al. 2012; Tuzzi et al. 2010). Both of them seem to be an effective tool of text classification with intelligible linguistic interpretation. As for the particular methods of analysis, secondary thematic concentration (*STC*), moving average type-token ratio (*MATTR*), and relative repeat rate (*RRMC*) are used in this study (for details, see below). This contribution follows up a similar research based on Czech data (Čech 2016).

 The basic assumption of this study is that thematic concentration and vocabulary richness are interdependent. More specifically, thematic concen­tration is based on so called thematic words (*TW*). *TW* are highly frequent autosemantics above *h*-point in the rank-frequency distribution of a text (see chapter 3.1). One can therefore assume that text with poor vocabulary should generate more words with high frequency and, consequently, more thematic words. In other words, we expect a significant negative correlation between vocabulary richness and thematic concentration.

Language Material

There are two corpora in this study. The first corpus (hereinafter C1) consists of English fiction texts, specifically 400 individual chapters of several novels writ­ten by Mark Twain, Jack London, Arnold Bennet, Charles Dickens, Henry James, and Thomas Hardy were chosen. In addition to these texts, we collected also the second corpus (hereinafter C2) which comprises 516 English texts of 6 genres (letter, news, poem, political speech, scientific text, short story) in order to discover whether genre can affect the assumed correlation between thematic concentration and vocabulary richness. It is worth mentioning that the corpora are not lemmatized. Thus, a wordform is a basic unit in this research. The part­icular methods (see Section 3) are applied to individual texts in both corpora. For text processing software *QUITA – Quan­titative Index text Analyzer* (Kubát et al. 2014) and *MaWaTaTaRaD* (Milička 2013) were used.

Methodology

Relative Repeat Rate (*RRMC*)

Repeat rate (*RR*) is a simple indicator of a degree of vocabulary concentration. In fact, *RR* measures vocabulary richness inversely: the higher *RR* is, the less vocabulary diversity a text has. *RR* is defined as follows:



*f1*…frequency of word *i* in a text

*N*…number of tokens

*V*…number of types

Figure 3. Relation between *RRMC* and text length (*N*) in 400 English texts (C1).

Results

In order to investigate the relation between vocabulary richness and thematic concentration in more detail, we compare several genres (letter, news, poem, political speech, scientific text, short story). The results of Kendall's tau correlation coefficient can be seen in Table 1. The obtained values mostly correspond to the previous ones. With the exception of poems, *MATTR* does not significantly correlates with *STC*, whereas *RRMC* significantly correlates with *STC* in 5 of 6 genres. Consequently, it can be concluded that genre probably does not substantially influence the relation between vocabulary richness and thematic concentration.

Correlations between *MATTR* and *STC*.

|  |  |  |  |
| --- | --- | --- | --- |
| genre | number of texts | *τ* | *p*-value |
| letter | 100 | 0.067 | 0.327 |
| news | 100 | -0.047 | 0.488 |
| poem | 100 | -0.196 | 0.028 |
| political speech | 56 | -0.154 | 0.375 |
| scientific text | 60 | -0.081 | 0.66 |
| short story | 100 | 0.040 | 0.063 |

Conclusion and Discussion

Considering the results, the final conclusion is quite ambiguous. Specifically, *RRMC* significantly positively correlates with *STC*, while *MATTR* seems to be independent on *STC*. Moreover, both results do not support our assumption, i.e. the negative correlation between vocabulary richness and thematic concentration. To sum up, this study raised more questions than answers.

We suppose that one of possible explanations could be the fact that *STC* is based on a relatively small number of thematic words (with regard to number of all types used in the text). The number of these frequent autosemantics above *h*-point is usually around 7 (but sometimes only 2 or 3; rarely even 0). Thus, it seems reasonable to assume that frequencies of these few words cannot significantly affect a resulting value of vocabulary richness measure which is based on frequencies of all words in a text. Needless to say, this idea must be scrutinized empirically. Further, the concept of vocabulary richness itself is still not clear and well theoretically based, despite decades of research. For instance, some authors consider *TTR* to be a matter of information flow rather than vocabulary richness (e.g. Popescu et al. 2009; Wimmer 2005). Until vocabulary richness is thoroughly and deeply examined, it will be very difficult and problematic to deal with this concept in stylometry.

From a point of view of this study, our preliminary findings must be especially verified by (a) an application of more vocabulary richness indices, and (b) more texts, particularly in different languages.

References

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

Čech, R., Garabík, R., Altmann, G. (2015). Testing the thematic concentration of text. *Journal of Quantitative Linguistics*, 22, 215–232.

Čech, R., Kubát, M. (2016). Text length and the thematic concentration of text. *Mathematical Linguistics*, 2(1), 5–13.