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

1 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 (RR_{MC}) 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.

2 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.

3 Methodology

3.1 Thematic Concentration

Every author of any text focuses on a topic or topics which are represented by several autosemantic words. Thematic concentration measures how intensively the author concentrates on the main theme(s) of the text. On the one hand, texts like scientific papers have usually high thematic concentration. On the other hand, e.g. informal letters or emails are not so thematically concentrated in general. There are several methods for measuring thematic concentration. In this study, we use secondary thematic concentration (STC) especially due to its effectiveness of text classification (Čech et al. 2015; Čech 2016).
STC is based on rank-frequency distribution and \( h \)-point (which represents a fuzzy boundary between synsemantic and autosemantic words; see formula 2). STC is calculated as follows:

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

- \( f(l) \)…highest frequency
- \( h \)…\( h \)-point
- \( r' \)…rank of an autosemantic word above \( h \)-point
- \( f(r') \)…frequency of autosemantic word

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

\( r \)…rank
\( f(r) \)…frequency of the rank

STC is considered to be the stylomeric index which is independent on text length (Čech, Kubát 2016). In Figure 1, the relation between the text length and STC in 400 English texts (C1) can be seen. Both, the low value of the coefficient of determination \( R^2 = 0.0016 \) and almost horizontal line of the linear function expressing the relationship between these indices can be considered as a sufficient support of STC text length independence.

![Figure 1. Relation between STC and text length (N) in 400 English texts (C1).](image)
3.2 Vocabulary Richness

Vocabulary richness is one of the most traditional stylometric features. In this study, we decide to use two indices: (a) moving average type-token ratio (\textit{MATTR}) and (b) relative repeat rate (\textit{RRMC}). These methods were chosen especially due to their strong resistance to the impact of the text length (Kubát 2014, McIntosh 1967).

3.2.1 Moving average type-token ratio (\textit{MATTR})

This vocabulary richness measure was proposed by Covington & McFall (2010) and further elaborated by Kubát & Milička (2013). \textit{MATTR} is defined as follows; a text is divided into overlapped subtexts of the same length (so called “windows” with arbitrarily chosen size \(L\); usually, the “window” moves forward one token at a time), next, type-token ratio is computed for every subtext and, finally, \textit{MATTR} is defined as a mean of the particular values. For example, in the following sequence of characters: \(a, b, c, a, a, d, f\), text length is 7 tokens \((N = 7)\) and we choose the window size to 3 tokens \((L = 3)\). We get 5 subsequent windows:

\[ |a, b, c| b, c, a| c, a, a| a, a, d| a, d, f|, \]

and compute \textit{MATTR} of the sequence as follows:

\[
\text{(3) } \text{MATTR}(L) = \frac{\sum_{i=1}^{N-L} V_i}{L(N - L + 1)} = \frac{3 + 3 + 2 + 2 + 3}{3(7 - 3 + 1)} = 0.87
\]

\(L\)...arbitrarily chosen length of a window, \(L < N\)
\(N\)...text length in tokens
\(V_i\)...number of types in an individual window

Although \textit{MATTR} was proposed as an absolutely independent method on text length, in Figure 2 we can observe a slight dependence in our corpus (C1) (coefficient of determination \(R^2=0.031\)). Nevertheless, \textit{MATTR} seems to be an appropriate index for the given purpose of this analysis.
3.2.2 Relative Repeat Rate ($RR_{MC}$)

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:

$$RR = \frac{1}{N^2} \sum_{i=1}^{V} f_i^2$$

$f_i$…frequency of word $i$ in a text  
$N$…number of tokens  
$V$…number of types

Given that the resulting values of $RR$ lie within the interval $<1/V;1>$, McIntosh (1967) proposed the relative repeat rate ($RR_{MC}$). Since the results of $RR_{MC}$ lie within the interval $<0;1>$, this relative repeat rate is comparable with other indicators such as $MATTR$. The formula is as follows:

$$RR_{MC} = \frac{1 - \sqrt{RR}}{1 - 1/\sqrt{V}}$$

![Figure 2. Relation between $MATTR$ and text length ($N$) in 400 English texts (C1).](image)
As can be seen in Figure 3, $RR_{MC}$ is not too much influenced by text length. The slight negative correlation with the coefficient of determination $R^2 = 0.031$ can be considered as an acceptable value for this research.

![Figure 3. Relation between $RR_{MC}$ and text length ($N$) in 400 English texts (C1).](image)

4 Results

As can be seen in Chyba! Nenalezen zdroj odkazu., MATTR seems to be independent on $STC$, the coefficient of determination $R^2 = 0.0007$. To be more precise, we apply also Kendall’s tau correlation coefficient with the results as follows: $\tau = -0.024$, $p = 0.466$. These results mean that there is non-significant ($\alpha = 0.05$) very slight negative correlation.
Contrary to MATTR, \( RR_{MC} \) significantly correlates with \( STC \), see Chyba! Nenalezen zdroj odkazů. \( (R^2 = 0.2151, \tau = 0.337, p < 0.001) \).

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 and Table 2. The obtained values mostly
correspond to the previous ones. With the exception of poems, $MATTR$ does not significantly correlates with $STC$, whereas $RR_{MC}$ 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.

Table 1

<table>
<thead>
<tr>
<th>genre</th>
<th>number of texts</th>
<th>$\tau$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>letter</td>
<td>100</td>
<td>0.067</td>
<td>0.327</td>
</tr>
<tr>
<td>news</td>
<td>100</td>
<td>-0.047</td>
<td>0.488</td>
</tr>
<tr>
<td>poem</td>
<td>100</td>
<td>-0.196</td>
<td>0.028</td>
</tr>
<tr>
<td>political speech</td>
<td>56</td>
<td>-0.154</td>
<td>0.375</td>
</tr>
<tr>
<td>scientific text</td>
<td>60</td>
<td>-0.081</td>
<td>0.66</td>
</tr>
<tr>
<td>short story</td>
<td>100</td>
<td>0.040</td>
<td>0.063</td>
</tr>
</tbody>
</table>

Table 2. Correlations between $RR_{MC}$ and $STC$.

<table>
<thead>
<tr>
<th>genre</th>
<th>number of texts</th>
<th>$\tau$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>letter</td>
<td>100</td>
<td>-0.020</td>
<td>0.77</td>
</tr>
<tr>
<td>news</td>
<td>100</td>
<td>0.193</td>
<td>0.004</td>
</tr>
<tr>
<td>poem</td>
<td>100</td>
<td>0.476</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>political speech</td>
<td>56</td>
<td>0.358</td>
<td>0.004</td>
</tr>
<tr>
<td>scientific text</td>
<td>60</td>
<td>0.261</td>
<td>0.008</td>
</tr>
<tr>
<td>short story</td>
<td>100</td>
<td>0.236</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Considering the obtained results, one can ask about a relation between $MATTR$ and $RR_{MC}$. A significant positive correlation can be seen in $Chyba! Nenalezen zdroj odkazů$. ($R^2 = 0.026$, $\tau = 0.091$, $p < 0.007$), however, the value of $\tau$ is very small.
Figure 6. Correlation between MATTR and RRMC in 400 English texts (C1).

5 Conclusion and Discussion

Considering the results, the final conclusion is quite ambiguous. Specifically, \( RR_{MC} \) 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

\begin{equation}
y = 0.0538x + 0.8913
\end{equation}

\( R^2 = 0.026 \)

\( RRMC \)

\( MATTR \)

\( 0.65 \quad 0.67 \quad 0.69 \quad 0.71 \quad 0.73 \quad 0.75 \quad 0.77 \quad 0.79 \)

\( 0.9 \quad 0.91 \quad 0.92 \quad 0.93 \quad 0.94 \quad 0.95 \)

It is worth mentioning that a number of thematic words correlates with text length. However, \( STC \) is not influenced by text length due to the normalization by dividing each thematic unit by the sum of all the weights of all the units above the \( h \)-point and the highest frequency of the unit in the text (see Formula 1). For example, a correlation between thematic words and text length in 57 political speeches is displayed in a graph in the appendix of this paper.
and deeply examined, it will be very difficult and problematic to deal with this concept in stylemetry.

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


---

2 Czech was analysed by Čech (2016) with similar results which generally correspond to our findings.
Appendix

Correlation between number of thematic words (TW) and text length (N) in 57 political speeches. [EK: please give for information about the data; it is not quite clear from where the data are coming]