# Vocabulary Richness Measure in Genres[1]

Miroslav Kubát[1] and Jiří Milička[1,2]

[1] Department of General Linguistics, Palacký University, Olomouc, Czech Republic

[2] Institute of Comparative Linguistics, Charles University, Prague, Czech Republic (milicka.cz)

## ABSTRACT

The paper deals with the one of the oldest and most traditional fields in quantitative linguistic, the concept of vocabulary richness. Although there are several methods for vocabulary richness measurement, all of them are influenced by text size. Therefore, the authors propose a new way of vocabulary richness measurement without any text length dependence. In the second part of the article, the new method is used for a genre analysis in texts written by the Czech writer Karel Čapek. There are also secondary analysed differences between authors and between languages.

Keywords: vocabulary richness, type-token ratio (TTR), stylometry, genre analysis, authorship attribution.

## 1. INTRODUCTION

Vocabulary richness measurement is one of the oldest and most traditional fields in quantitative linguistics. The concept of vocabulary richness measure is based on the fact that each person uses a specific individual vocabulary. Linguists use the concept of vocabulary richness mostly in authorship and genre analysis. One of the oldest and easiest ways of vocabulary richness measure is the type-token ratio (TTR). The TTR index is based on the simple ratio between the number of types and tokens in a text. The resulting value shows how much the vocabulary varies (the more vocabulary variation in a text, the higher TTR).

The stumbling block of TTR and all indexes based on word frequency is the fact that there is a dependence on text size. Although many attempts to reduce this problem were proposed, no one was fully successful (most notable in recent years $R_1$ and Lambda structures proposed by

[1] This is an Author's Original Manuscript of an article submitted for consideration in the Journal of quantitative linguistics © Taylor & Francis; the published version is available online at http://www.tandfonline.com/10.1080/09296174.2013.830552*].*

Popescu et al. 2009, 2011). Another disadvantage of indexes measuring vocabulary richness is the fact that the result is mostly only one figure, which can be misleading. One of the most comprehensive books giving an overview in this field is Word frequency studies (Popescu et al. 2009). Given that all proposed formulas failed, it is necessary to find a new solution.

Since vocabulary richness is mostly used in stylometry, we analysed genres in texts written by the Czech author Karel Čapek. We decided to use a corpus consisting of texts written by only one author to avoid a bias caused by different authors' styles. The main aim of the analysis is to discover whether we can distinguish genres using this feature. We follow up the work of Marie Těšitelová who established the usage of statistical methods in Czech linguistics and brought several studies in this field (e.g. Těšitelová 1974, 1983, 1987).

This research has two aims. The first one is to propose a new way of vocabulary richness measure without any text size dependence. The second one is to discover whether vocabulary richness is an advisable criterion for genre attribution.

## 2. DEFINITIONS

### 2.1 Moving Average Type-Token Ratio (MATTR)

Considering the dependence between the text length and plain type-token ratio, Moving Average Type-Token Ratio (MATTR) was proposed by Covington and McFall (2010, p. 96-97). The definition is as follows (freely quoted):

Consider a text consisting of words $w_1$ to $w_n$ and number L arbitrarily chosen where L<N, where N denotes the length of the given text in term of running words.

For each i; $i \in \mathbb{N}$, $i < N - L$ iterate following two steps:

1. Select the subtext $w_i$ to $w_{i+L}$.
2. Count the number of types ($V_i$) in the subtext.

The average type token ration $MATTR(L)$ is defined as:

$$MATTR(L) = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L)}$$

The main disadvantage of the MATTR is that it produces only one figure (e.g. the novel Krakatit written by Karel Čapek has $MATTR(100) = 0.78$), which may result in misleading interpretations when comparing the measure of one text with another one.

The idea of a moving window is not new; it is implemented in the software WordSmith (Scott, M., 2013) as the standardized type-token ratio (STTR) where the average TTR is based

on consecutive word chunks of a text; STTR is based on non-overlapping windows whereas MATTR uses smoothly moving window.

### 2.2 Moving Window Type-Token Ratio (MWTTR)

Moving Window Type-Token Ratio can be defined as the series of $V_i$ (or by another words, each $V_i$ is mapped to its i ). An example follows:

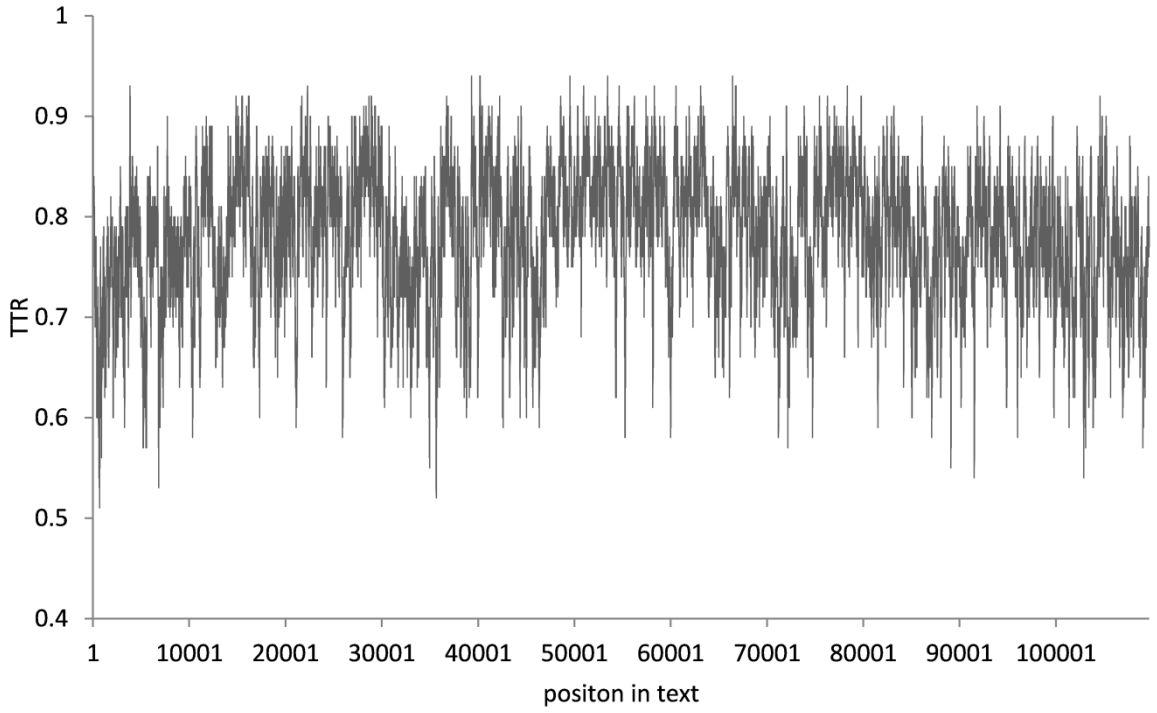


Figure 1. Results of MWTTR(100) in the novel Krakatit

The MWTTR has been proposed by Reinhard Köhler and Matthias Galle (1993) (although not called MWTTR) and it was used also by Covington and McFall (2010, p. 98) (albeit not defined nor called MWTTR).

### 2.3 Moving Window Type-Token Ratio Distribution (MWTTRD)

The MWTTR is suitable to study changes of the TTR value within one text, but is not appropriate to study the TTR of the text as a whole. Thus we propose Moving Window Type-Token Ratio Distribution – the distribution of MWTTR values. By terms of the previous subsections: to each $a_{j;}, j \in \mathbb{N}, j \leq L$ map the number of the iterations in which $V_i$ =j.

The usage of the method is illustrated in the following chart:

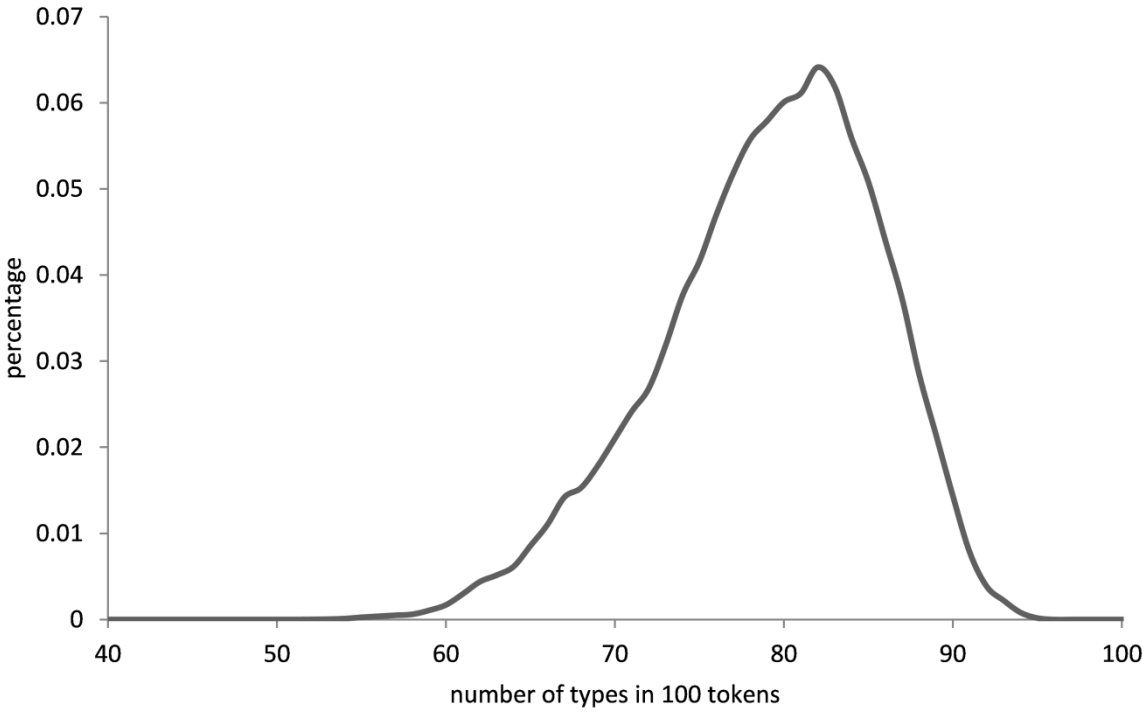


Figure 2. Results MWTTRD(100) in the novel Krakatit

In Figures 1 and 2 can be seen that MWTTR focuses on a development in a text whereas our measurement considers a text as a whole.

The method was implemented in the MaWaTaTaRaD freeware.[2]

## 3. METHODOLOGY

The word-forms are used as units for all calculations in this research. Thus, no text was lemmatized. The main reason for this decision lies in the fact that there is not general consensus how to lemmatize text and the word-form segmentation is thus less ambiguous. Moreover, this method allows comparing results obtained from analyses in different languages.

The cornerstone of every quantitative analysis is an appropriate sample. Given that we aim to discover possible differences between genres, the sample contains texts written by only one author. This method secures results from negative influence of different authors' styles. We chose texts written by the Czech author Karel Čapek who wrote many texts in several genres.

[2] Available on http://www.milicka.cz/mawatatarad. MATTR and MWTTR are also included in the software.

We matched up his texts with seven genres (travel book, novel, short story, children's literature, correspondence, scientific text, poem).

MWTTRD(100) was computed for the mentioned texts and the results were compared. In this research, we chose L=100 for all calculations.

## 4. RESULTS

The resulting values of each genre can be seen in Figure 3.

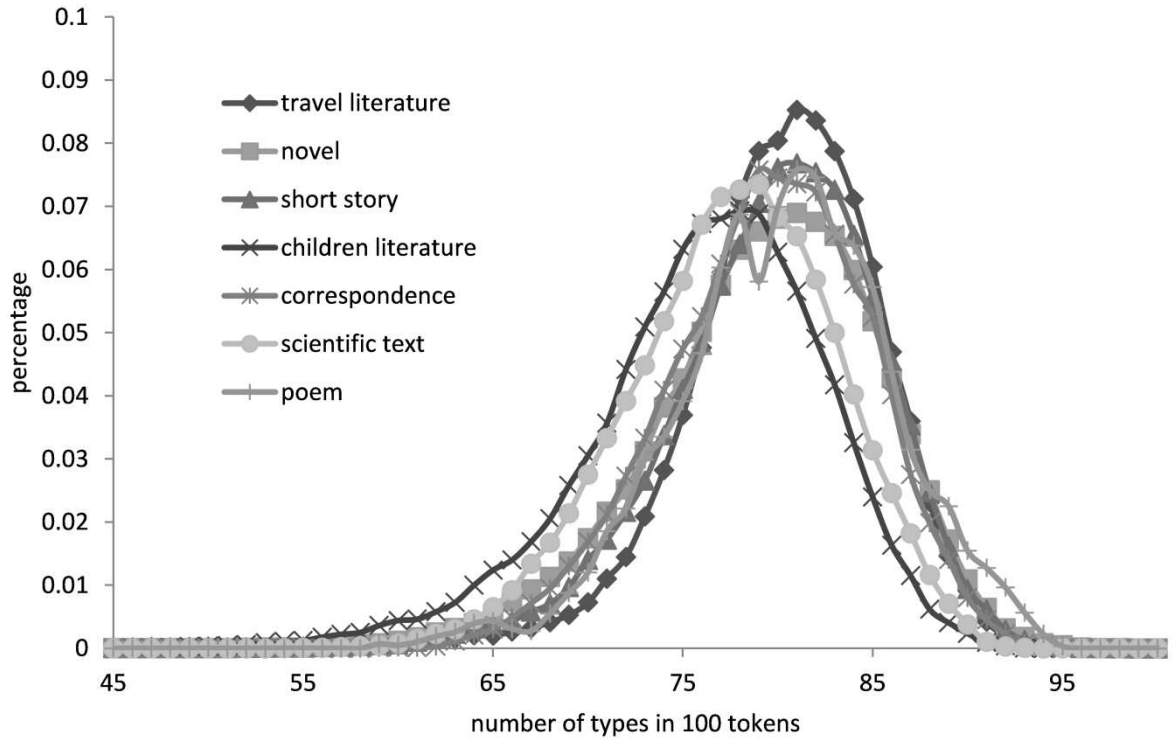


Figure 3. TTR in genres

Although the curves seem to be very similar, we must discover the differences between the curves in a more proper way. We decided to use the so called $\chi^2$ discrepancy coefficient (C) (see Mačutek 2013) which is usually used for the measurement of goodness of fit. We consider value C = 0.05 to be a limit for the decision whether two distributions are similar or not (the lower C, the more similar distributions are). The results of the discrepancy coefficient can be seen in Table 1.

Table 1. Results of the discrepancy coefficient in genres (values C ≥ 0.05are highlighted in bold)

| | travel book | novel | short story | children's literature | correspondence | scientific text | poem |
|---|---|---|---|---|---|---|---|
| travel book | x | | | | | | |
| novel | 0.013 | x | | | | | |
| short story | 0.006 | 0.004 | x | | | | |
| children's literature | **0.128** | 0.035 | **0.071** | x | | | |
| correspondence | 0.017 | 0.002 | 0.004 | **0.074** | x | | |
| scientific text | **0.076** | 0.018 | 0.039 | 0.020 | 0.027 | x | |
| poem | 0.005 | 0.001 | 0.001 | **0.051** | 0.009 | 0.029 | x |

Considering the results in Table 1, we can say that TTR is not a very suitable tool for distinguishing differences between genres. Nevertheless, we discovered the extraordinary position of children's literature between genres. This genre differs from four of six other ones. We assume that this fact is caused by the need for a limited vocabulary due to readability for children. Although one can expect also an extraordinary position of poems, the results reject such expectations.

Since vocabulary richness seems to be not very powerful for genre analysis, one can ask whether we can use the measurement for authorship attribution. Therefore, we compared eight Czech authors (namely K. Čapek, A. Jirásek, F. L. Čelakovský, K. Havlíček, K. J. Erben, O. Březina, S. Čech, V. Vančura) using the same method. The corpora consist of more than sixty books. The curves of the TTR distributions can be seen in Figure 4.

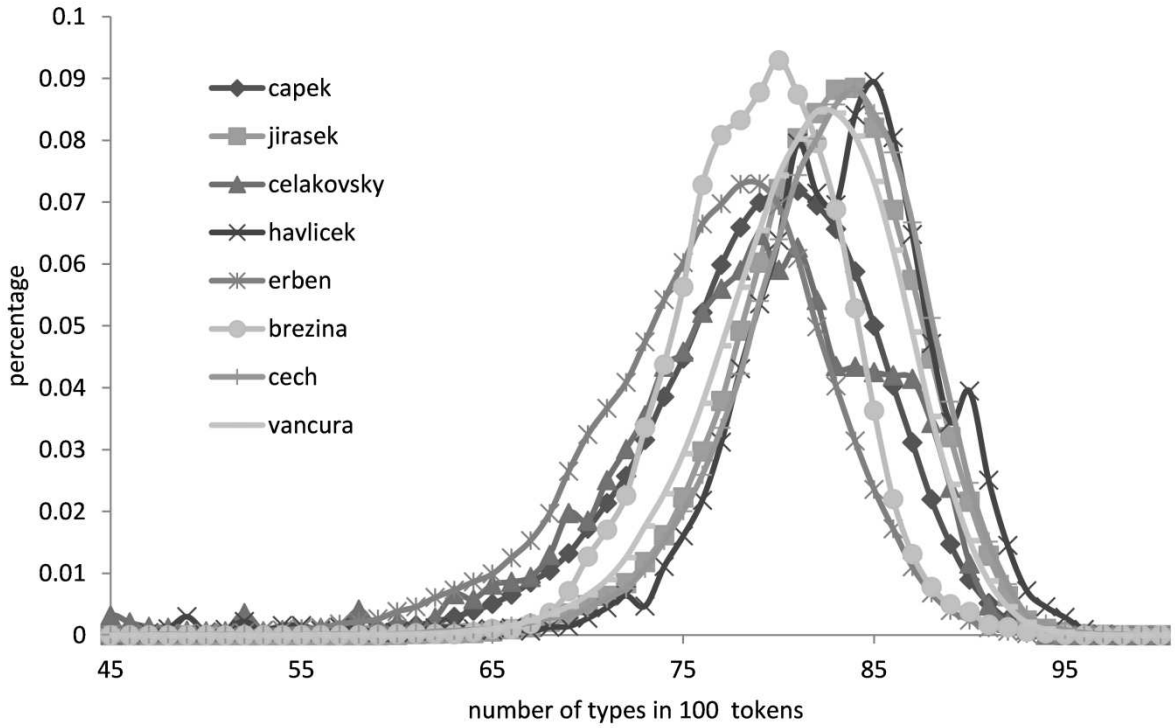


Figure 4. TTR in authorship

At first sight, the differences between the authors in Figure 4 seem to be greater than between the genres in Figure 3. The results of the discrepancy coefficient are shown in Table 2.

Table 2. Results of the discrepancy coefficient in authorship (values C ≥ 0.05are highlighted in bold)

| | capek | jirasek | celakovsky | havlicek | erben | brezina | cech | vancura |
|---|---|---|---|---|---|---|---|---|
| capek | x | | | | | | | |
| jirasek | 0.044 | x | | | | | | |
| celakovsky | 0.009 | **0.073** | x | | | | | |
| havlicek | 0.006 | 0.016 | **0.119** | x | | | | |
| erben | 0.030 | **0.243** | 0.037 | **0.095** | x | | | |
| brezina | 0.005 | **0.068** | **0.092** | **0.156** | 0.038 | x | | |
| cech | 0.043 | 0.004 | **0.080** | 0.009 | **0.265** | **0.110** | x | |
| vancura | 0.044 | 0.004 | 0.030 | 0.013 | **0.126** | 0.012 | 0.009 | x |

According to the discrepancy coefficient values in Table 2, it is evident that vocabulary richness is a quite appropriate feature for authorship analysis. For a better clarity, Figure 5 shows a network where the authors with similar MWTTRD are connected.

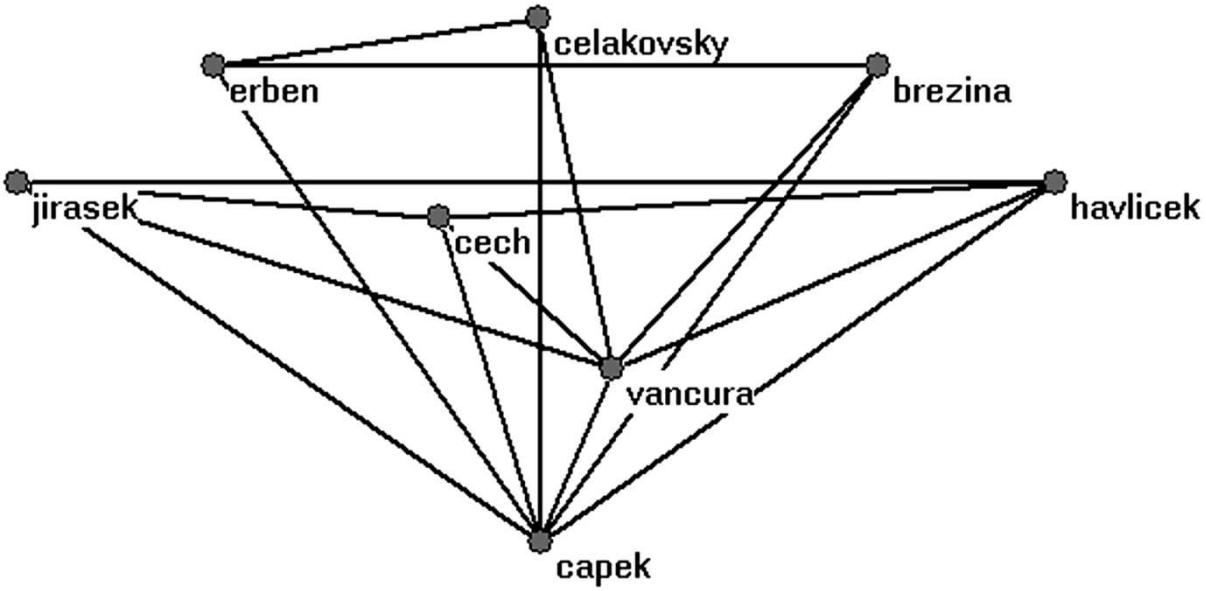


Figure 5. The network in which the authors with similar MWTTRD are connected

Based on the results in Table 2 and Figure 5, we can state that three authors (Čelakovský, Erben, Březina) have an extraordinary position between the eight analysed writers. Březina's poetry belongs to symbolism, his writing is full of metaphors, philosophical and scientific terms. Therefore, his poems aimed to a small circle of intellectual readers. In contrast to Březina; Čelakovský and Erben wrote folk poetry based on oral texts. The style of these texts is simple and is connected to less vocabulary richness. Although one can expect the extraordinary position of these writers, it is quite surprising that Březina does not differ from Erben. Considering the aforementioned short literary background, we can state that TTR measurement is a more or less suitable method for authorship analysis.

Since we applied the new method of vocabulary richness measure to genre and authorship analysis, it is logical to ask whether the measurement can be used for distinguishing languages. Therefore, we created a corpus consisting of eight languages with different typology (namely Czech, German, Italian, Hungarian, Arabic, Tagalog, English and Basque). To obtain comparable results, each language is represented by 10 long prosaic texts. The results are displayed in Figure 6.

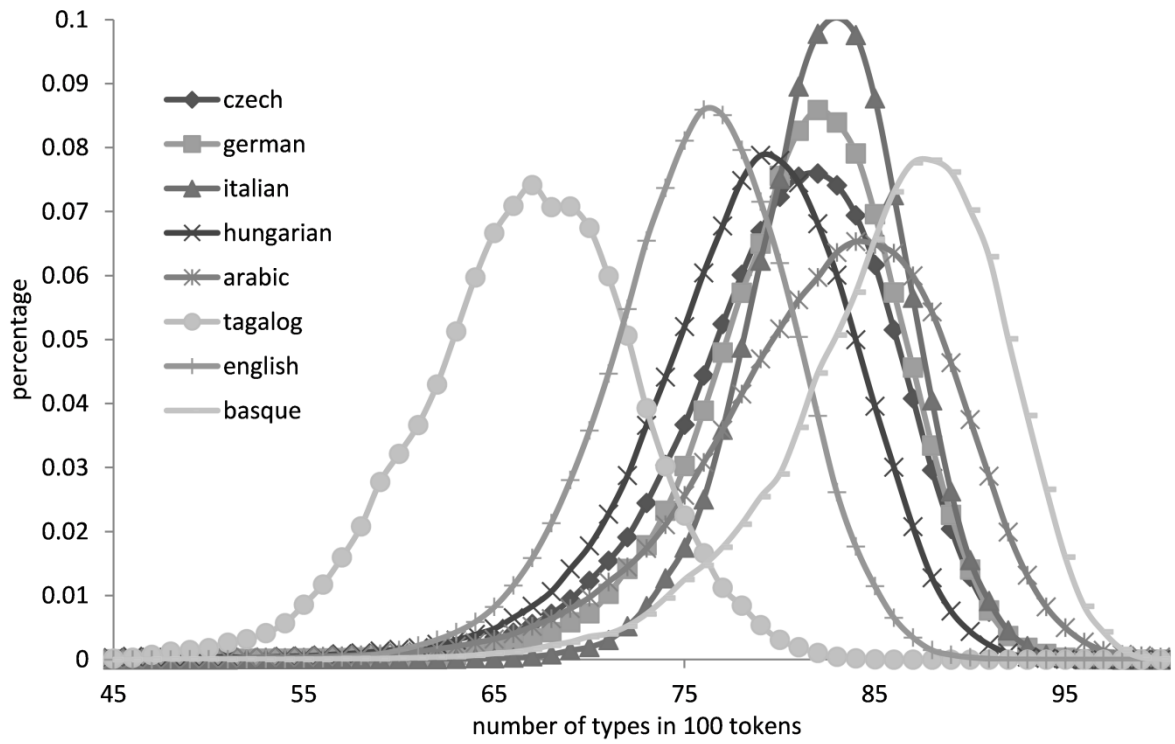


Figure 6. TTR in languages

The results of the discrepancy coefficient can be seen in Table 3:

Table 3. Results of the discrepancy coefficient in languages (values C ≥ 0.05are highlighted in bold)

| | czech | german | italian | hungarian | arabic | tagalog | english | basque |
|---|---|---|---|---|---|---|---|---|
| czech | x | | | | | | | |
| german | 0.004 | x | | | | | | |
| italian | 0.023 | 0.026 | x | | | | | |
| hungarian | 0.019 | **0.066** | **0.154** | x | | | | |
| arabic | **0.057** | 0.038 | **0.057** | **0.104** | x | | | |
| tagalog | **0.336** | **0.678** | **0.821** | **0.485** | **0.472** | x | | |
| english | **0.119** | **0.263** | **0.386** | **0.091** | **0.277** | **0.359** | x | |
| basque | **0.154** | **0.218** | **0.197** | **0.369** | **0.050** | **0.826** | **0.570** | x |

The discrepancy coefficient values in Table 3 are high, when comparing to the previous ones. The languages are similar only in five of 28 cases. It is interesting that in Figure 6 distances between languages seem to be correlated with geographical location rather than with the typological differences. Given that this research is not primarily aimed to language analysis, we will not deal with this issue in detail. Nevertheless, it could be a remarkable observation for future language researches. In our context, it is primarily important that we can consider MWTTRD to be a very powerful tool for language analysis.

## 5. CONCLUSION AND DISCUSSION

This work consists of two main parts, the first one is the new method of vocabulary richness measurement, the second one is genre analysis based on the proposed method.

We proposed this new method of vocabulary measurement (Moving Window Type-Token Ratio Distribution; MWTTRD) which is independent on text length. In contrast to other methods, we consider the entire distribution in the measurement. Therefore our method can be used for the analysis of texts with different lengths and the results are not limited by only one resulting value.

The research also brought several important observations. Vocabulary richness measurement seems to be not very efficient tool for genre analysis. We discovered that only one genre (children's literature) has an extraordinary position. This genre differs from four of six other ones. On the other hand, we analysed only texts written by one Czech author, therefore it is necessary to analyse more texts from other authors and languages. According to our results in authorship analysis, we consider vocabulary richness to be a matter of authorship rather than genre. However, the best results were obtained in language analysis where almost all languages were mutually different.

Finally, it must be said that this work is just a first attempt to discover whether vocabulary richness is a suitable feature for genre analysis. Therefore, it is necessary to analyse more texts to support or reject our preliminary claims.

## ACKNOWLEDGEMENTS

We would like to thank Reinhard Köhler for helpful comments and suggestions. We are also grateful to Ján Mačutek for his help with statistics. This work was supported by the project Lingvistická a lexikostatistická analýza ve spolupráci lingvistiky, matematiky, biologie a psychologie, grant no. CZ.1.07/2.3.00/20.0161 which is financed by the European Social Fund and the National Budget of the Czech Republic.

## REFERENCES

Altmann, G., Wimmer, G. (1999). Review Article: On Vocabulary Richness. Journal of Quantitative Linguistics, 6(2), 1–9.

Čech, R. (2011). Frequency structure of New Year's presidential speeches in Czech. The authorship analysis. In E. Kelih et al. (eds) Issues in Quantitative Linguistics 2. Lüdenscheid: RAM-Verlag, 82–94.

Covington, M. A., McFall J. D. (2008). The Moving-Average Type-Token Ratio. Presented as a poster at the Annual Meeting of the Linguistic Society of America.

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.

García A. M., Martín, J. C. (2005). The validity of lemma-based lexical richness in authorship attribution: A proposal for the Old English Gospels. ICAME Journal, 29, 115–130.

Hoover, D. L. (2003). Another Perspective on Vocabulary Richness. Computers and the Humanities, 37, 151–178.

Jamak, A., Savatić, A., Can, M. (2012). Principal component analysis for authorship attribution. Business Systems Research, 3(2), 49–56.

Köhler, R., Gale,M. (1993): Dynamic Aspects of Text Characteristics. In L.Hřebíček, G. Altmann (eds.) Quantitative Text Analysis. Trier, Wissenschaftlicher Verlag, 46-53.

Mačutek, J (2013): Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics; forthcoming JQL 2013

Mikros, G. K., Argiri, E. K. (2007). Investigating topic influence in authorship attribution. In B. Stein, M. Koppel & E. Stamatatos (eds), Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, 276, 29–35. Amsterdam, Netherlands: CEUR

Milicka, J. (2013). MaWaTaTaRaD. Prague. (Software)

Mistrík, J. (1969). Frekvencia slov v *slovenčině*. Bratislava: Slovenská akadémia vied.

Mistrík, J. (1985). *Frekvencia tvarov a konštrukcií v slovenčině*. Bratislava: Veda.

Mistrík, J. (1989). *Štylistika*. Bratislava: SPN.

Peng, R. D. , Hengartner, N. W. (2002). Quantitative Analysis of Literary Styles. The American Statistician, 56(3), 175–185.

Popescu, I.-I. et al. (2009). Word frequency studies. Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. Glotometrics, 13, 23–46.

Popescu, I.-I., Altmann, G., Čech, R. (2011). The Lambda-structure of Texts. Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). Aspects of Word Frequencies. Lüdenscheid: RAM-Verlag.

Rohangiz M. D. (2007). Authorship Attribution and Statistical Text Analysis. *Metodološki zvezki*, 4(2), 149–163.

Scott, M. (2013). WordSmith Tools. Liverpool: Lexical Analysis Software.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

Stamatatos, E., Fakotakis, N., Kokkinakis, G. (2001). Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics 26(4), 471–495.

Těšitelová, M. (1974). *Otázky lexikální statistiky*. Praha: Academia.

Těšitelová, M. (1987). *Kvantitativní lingvistika*. Praha: SPN.

Těšítelová, M. et al. (1983). *Psaná a mluvená odborná čeština z kvantitativního hlediska*. Praha: Ústav pro jazyk český ČSAV.

Těšitelová, M. et al. (1987). *O češtině v číslech*. Praha: Academia.

Wimmer, G. (2005). The type-token relation. In R. Köhler, G. Altmann, R. G. Piotrowski (eds) Quantitative Linguistics. An International Handbook. Berlin: de Gruyter. 361–368.

Wimmer, G. et al. (2003). *Úvod do analýzy textov*. Bratislava: Veda.