Chapter 4
Morphological Richness of Text

Radek Čech and Miroslav Kubát

Abstract This study proposes a method for measuring the morphological richness of text. The method enables us to characterize the morphological complexity of a text (or a corpus). It is based on a computation of the difference between two measurements — the vocabulary richness of lemmas and the vocabulary richness of word forms. The greater the difference, the higher the morphological complexity of a text. The Moving Average Type Token Ratio (*MATTR*) is used for the computation of vocabulary richness. We hypothesize that the proposed indicator, known as Moving Average Morphological Richness (*MAMR*), should reflect the style of a text, and could therefore be used in stylometry. To verify this assumption, *MAMR* is applied in analyses of both genre and authorship.

Keywords Morphological richness · Vocabulary richness · Stylometry · Genre · Authorship · Czech language

Introduction

Any text can be seen as the result of miscellaneous factors. A writer (or a speaker) has many different choices to apply his or her language competence. Furthermore, it is obvious that humans use these choices intensively. Take a group of people with the same age, educational background, sex, and IQ and ask them to write a text focused on the same topic in a very specific genre; there will be just a few identical clauses (if any) and no identical paragraphs (e.g., Cvrček & Václavík, 2015; Indrisano & Squire, 2000; Pinker, 2010).[1] This well-known fact, i.e., the huge degree of variability in language use, has been recognized among linguists for many decades, and it represents a fundamental condition for any analysis of style and authorship (e.g., Juola, 2008; Kubát, 2016). There are many properties of a text that

[1] It should be pointed out that this issue is beyond the scope of this study.

R. Čech (✉) · M. Kubát
University of Ostrava, Ostrava, Czech Republic

M. Fidler, V. Cvrček (eds.), *Taming the Corpus*, Quantitative Methods in the Humanities and Social Sciences, https://doi.org/10.1007/978-3-319-98017-1_4

reflect its uniqueness, and some of these properties are more "visible" than others. For instance, vocabulary richness seems to be an intuitively comprehensible and relatively easily observable property for comparing texts; similarly, the distribution of parts of speech could also be characterized as a "visible" property. By contrast, some abstract properties based on the so-called frequency structure of a text, such as lambda structure (Popescu, Čech, & Altmann, 2011) and the writer's view (Popescu & Altmann, 2007), are less "visible."

In this study, we introduce morphological complexity as a stylometric indicator, which can be applied to classify texts; we focus particularly on genre and authorship analysis. The concept of morphological complexity is widely used in language typology, and it has been investigated many times using various measurements (cf. Baerman, Brown, & Corbett, 2015; Bane, 2008; Bentz, Ruzsics, Koplenig, & Samardžić, 2016). It has also been applied in several other fields, such as child language acquisition or second language acquisition (cf. Březina & Pallotti, 2016; Xanthos et al., 2011). The advantage of this concept lies in its intelligible interpretation and relatively simple operationalization. However, its use in stylometry faces several problems that are typical for this kind of analysis; primarily, text length impact has to be eliminated to avoid misinterpretation of the results.

This study has two aims: (1) to propose a method for measuring the morphological complexity of texts, and (2) to observe whether this method is an effective tool for stylometric research. Thus, it should be emphasized that the aim of both the genre analysis and the authorship analysis in this study is to conduct a preliminary test measurement of morphological complexity in terms of text classification. The corpus was created in accordance with the aim of this study. We do not therefore analyze these texts from a literary perspective. The method is based on a computation of the difference between two measurements — the vocabulary richness of lemmas and the vocabulary richness of word forms. The greater the difference, the higher the morphological complexity of a text. For example, let us take two sentences that both consist of 10 tokens: "I was ready to be a member of the team" (S1) and "I was ready to become a member of the team" (S2). After lemmatization, the sentences would be "I be ready to be a member of the team" (S3) and "I be ready to become a member of the team" (S4). Further, for the measurement of vocabulary richness, Type-Token Ratio (*TTR*) is used here:

$$TTR_{S1} = \frac{10}{10} = 1$$

$$TTR_{S3} = \frac{9}{10} = 0.9$$

$$TTR_{S2} = \frac{10}{10} = 1$$

$$TTR_{S4} = \frac{10}{10} = 1$$

With sentences S1 and S3, we get a morphological complexity $TTR_{S1} - TTR_{S3} = 1 - 0.9 = 0.1$; whereas with sentences S2 and S4, the result of the morphological complexity is $TTR_{S2} - TTR_{S4} = 1 - 1 = 0$. Thus, we can state that S1 has higher morphological complexity than S2.

Since the Moving Average Type-Token Ratio (Covington & McFall, 2010; Kubát & Milička, 2013) is used for the measurement of vocabulary richness, the method is named the Moving Average Morphological Richness (hereinafter *MAMR*).

Using vocabulary richness for measuring the morphological complexity of a text is not new in linguistics; Kettunen (2014) applied the Moving Average Type-Token Ratio (hereinafter *MATTR*) directly in a cross-linguistic comparison (though not as a difference computation between word forms and lemmas). He computed *MATTR* for texts in 21 languages, and the results were compared with two other methods of measuring morphological complexity. The author states that "All the three computed measures are able to order the languages quite meaningfully in a morphological complexity order that at least groups most of the languages with same kind of languages and the most and least complex languages are clearly separated" (Kettunen, 2014). However, this approach seems to be problematic, because *MATTR* represents more than just morphological richness. Perhaps Kettunen's approach is acceptable in language typology, but in our opinion it is not suitable for stylometric research.

The morphological complexity of a text seems to be the result of unconscious language behavior by the writer (or the speaker); it is hard to imagine that the author of a text consisting of perhaps thousands of words consciously distributes the proportions of particular word forms. Moreover, the distribution of word forms is strongly influenced by grammar; the author is therefore "forced" to use particular forms regardless of his or her preferences. Consequently, no one can be sure that the concept of morphological complexity is useful for determining style or authorship attribution until this is empirically proved. Thus, one goal of this study is to observe whether the *MAMR* of a text can distinguish an individual style of writing — like other stylometric indicators such as thematic concentration (Čech, 2016), vocabulary richness, or activity of text (Kubát, Matlach, & Čech, 2014; Popescu et al., 2009). A corpus of 677 Czech texts written by eight authors is used for the analysis.

This paper is structured as follows. First, the methodology is introduced (section "Methodology"). In section "Corpus," the language material is presented and analyzed. Section "Text Length" is concentred to an observation of a potential impact of text length on all indices used in the current study. Section "Results" is devoted to the results, and "Conclusion" presents the conclusions of the study.

Methodology

The method of measuring morphological richness is based on a computation of the difference between the vocabulary richness of lemmas and the vocabulary richness of word forms. As "Introduction" illustrated, the bigger the difference, the higher the morphological complexity of the text.

Another set of examples is shown below. Let us take two seven-word texts as an example:

(1a) I love her and she loves me
(2a) I love it and she loves it

We lemmatize the texts as follows:

(1b) I LOVE SHE AND SHE LOVE I
(2b) I LOVE IT AND SHE LOVE IT

Since both texts are of identical length, it is possible to use the Type-Token Ratio (*TTR*) as an indicator of vocabulary richness:

$$TTR = \frac{V}{N}$$

where V is the number of different words (types) in a text and N is the number of all words (tokens) in a text. We compute the *TTR* for each text:

$$TTR_{1a} = \frac{7}{7} = 1$$

$$TTR_{2a} = \frac{6}{7} = 0.857$$

$$TTR_{1b} = \frac{4}{7} = 0.571$$

$$TTR_{2b} = \frac{5}{7} = 0.714$$

The difference between *TTRs* based on word forms and lemmas expresses the morphological complexity of a text; specifically, for text (1) we obtain:

$$TTR_{1a} - TTR_{1b} = 1 - 0.571 = 0.429$$

and for text (2)

$$TTR_{2a} - TTR_{2b} = 0.857 - 0.714 = 0.143$$

Since 0.429 > 0.143, one can state that text (1) has higher morphological complexity.

In reality, we need to compare texts of different lengths. Thus, the Moving Average Type-Token Ratio (hereinafter *MATTR*) for measuring vocabulary richness is applied because of its independence from text length (Covington & McFall, 2010;

Kubát & Milička, 2013).[2,3] *MATTR* is defined as follows. A text is divided into overlapping subtexts of the same length (so-called "windows" with arbitrarily chosen size *L*; usually, the "window" moves forward one token at a time). Then, the type-token ratio is computed for every single subtext, and finally *MATTR* is defined as a mean of the individual values:

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

where N is the text length in tokens, L is the arbitrarily chosen length of a window ($L < N$), and V_i is the number of types in an individual window.

For example, in the following sequence of characters — *a, b, c, a, a, d, f* — the text length is 7 tokens ($N = 7$). If we choose a window size of 3 tokens ($L = 3$), we obtain 5 windows — *a, b, c*|*b, c, a*|*c, a, a*|*a, a, d*|*a, d, f* — and then we can compute the *MATTR* of the sequence as follows:

$$MATTR(L) = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)} = \frac{3+3+2+2+3}{3(7-3+1)} = 0.867$$

The *MAMR* of a text is defined as the difference between the *MATTR* computed in word forms and the *MATTR* computed in lemmas:

$$MAMR(L) = MATTR(L)_{wordform} - MATTR(L)_{lemma}$$

Unfortunately, the nature of the measurement does not allow us to test differences between pairs of texts statistically.[4] However, it is possible to test differences between text groups (genres, authors). In this analysis, we use the *u*-test[5]:

$$u = \frac{\left|\overline{MAMR}_1 - \overline{MAMR}_2\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

[2] MATTR is a similar method to Standardized Type-Token Ratio (STTR). MATTR is based on overlapping chunks, while STTR is based on nonoverlapping chunks.

[3] Although MATTR is independent of text length, it should be mentioned that this method is problematic because of the arithmetic mean value of the chunks. For example, although two nonoverlapping text chunks (subtexts) can share the same TTR value, the inventory of types in these two chunks can be completely different. Another problem may arise when TTR on different chunks of text has a high variance. The authors of this study are aware of these problems, especially the high variance. That is why, the MWTTRD method was proposed (Kubát & Milička, 2013). Nevertheless, according to data obtained in the previous research (Kubát, 2016), the average value seems to be a reliable indicator for stylometric analyses.

[4] To put it more specifically, the problem is caused by overlapping windows.

[5] In statistics, it is usually called the *z*-test; here, we follow a convention used in quantitative linguistics.

where $\overline{MAMR}_1$ and $\overline{MAMR}_2$ are the arithmetic means of the results in each group, s_1, s_2 are standard deviations, and n_1, n_2 are the numbers of results in each group. For the significance level $\alpha = 0.05$, $u \geq 1.96$ means that the difference between the two groups is statistically significant.

For illustration, let us compare differences in *MAMR* between Karel Čapek's short stories (*Wayside Crosses* (WC), *Stories from a Pocket* (SP), *Stories from Another Pocket* (SAP), and *Painful Tales* (PT)[6]) on the one hand, and his newspaper articles (*How it is Made* (HM), *the Gardener's Year* (GY), and selected articles from The People's Newspaper (PN))[7] on the other. Using the data (texts WC, SP, SAP, PT, HM, GY, and PN), we obtain:

$$u = \frac{\left|\overline{MAMR}_{short\ stories} - \overline{MAMR}_{newspapers}\right|}{\sqrt{\frac{s^2_{short\ stories}}{n_{short\ stories}} + \frac{s^2_{newspapers}}{n_{newspapers}}}} = \frac{|0.0984 - 0.0797|}{\sqrt{\frac{0.0164^2}{71} + \frac{0.0216^2}{92}}} = 6.28$$

Since $6.28 > 1.96$, we can state that there is a significant difference between these two groups of texts (for the $\alpha = 0.05$).

Corpus

The proposed method is applied to a corpus of 677 Czech texts. For genre analysis, we decided to use texts only written by one author (Karel Čapek) in order to avoid biased results caused by different authorial styles. The texts belong to five genres: travel books (travelogues), letters, short stories, novels, and newspaper articles. However, it should be emphasized that such an analysis is limited to one particular author; we cannot generalize the findings to other authors, and the interpretation must take this fact into account. To carry out a more thorough genre analysis, texts by many authors must be investigated. The primary purpose of this study is to propose the method, and its secondary purpose is to conduct a preliminary test to discover whether *MAMR* has some potential for application in stylometric research. In other words, this article focuses on the method from the perspective of quantitative linguistics; it is not a literary genre analysis.

For the authorship analysis, novels written by eight Czech writers were chosen: Karel Čapek (1890–1938), Alois Jirásek (1851–1930), Božena Němcová (1820–1862), Vladislav Vančura (1891–1942), Bohumil Hrabal (1914–1997), Karel Poláček (1892–1945), and Svatopluk Čech (1846–1908). As in the case of the author-specific genre analysis material mentioned above (Čapek texts), this corpus too is used only for preliminary testing to assess *MAMR's* potential for authorship

[6] The Czech original titles: *Boží muka* (WC), *Povídky z jedné kapsy* (SP), *Povídky z druhé kapsy* (SAP), and *Trapné povídky* (PT).

[7] The Czech original titles: *Jak se co dělá* (HM), *Zahradníkův rok* (GY), and *vybrané články z Lidových novin* (PN).

attribution; the study does not present any literary interpretation of the results obtained.

For the purposes of this study, novels and travel books were segmented into individual chapters. Analogically, collections of short stories were segmented into individual short stories. In short, the following units were considered to be individual texts for the purposes of the present study: individual chapters of a novel or a travel book, and individual short stories, letters, and newspaper articles. The list of texts used for the genre and authorship analysis can be found in Appendix.

Text Length

Text length is a factor that influences the majority of indices used in stylometry. Needless to say, the impact of text length is undesirable, and researchers usually attempt to find some methods to eliminate it. Let us briefly mention other text size-independent methods based on *TTR*.

The idea of a moving window is not new; it is implemented in the software WordSmith (Scott, 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 windows.

Another standardized Type-Token Ratio, *zTTR*, was proposed by Cvrček and Chlumská (2015). This vocabulary richness indicator is based on comparing observed *TTR* with referential *TTR* values representing texts of identical size. The main disadvantage of *zTTR* is that it is based on a corpus which cannot be considered fully representative. The crucial question is how to select particular texts, e.g., a representative corpus of novels. There is no clear standard for selecting appropriate novels for the corpus.

Besides the aforementioned indicators, there are several other methods such as Moving Window Type-Token Ratio Distribution (*MWTTRD*) (Kubát & Milička, 2013), *R1* based on h-point (Popescu et al., 2009), a complex frequency structure indicator called lambda (Popescu et al., 2011), Yule's *K* (Yule, 1944), and Guiraud's *TTR* (Guiraud, 1954). All these methods have advantages and disadvantages; some are not fully independent of the text length, while some require specific text lengths.

The application of the "moving window" (see the *MATTR* in "Methodology") seems to be a promising method for eliminating the impact of text length. *MATTR's* advantage is in its straightforward interpretation and low computational complexity. On the other hand, this method has also some weaknesses (discussed above in this study). Nevertheless, according to data observations in the previous research (Kubát, 2016; Kubát & Milička, 2013), the *MATTR* seems to be a reliable indicator for stylometric analyses.

In this analysis, we observe the potential impact of text length on all indices used in the current study (i.e., $MATTR_{word\ form}$, $MATTR_{lemma}$, *MAMR*) (Figs. 4.1, 4.2, and 4.3). We decided to present these graphs, because text length is one of the most frequent obstacles to the use of stylometric indicators (especially, those related to vocabulary richness). In all cases, the variables are obviously independent of one

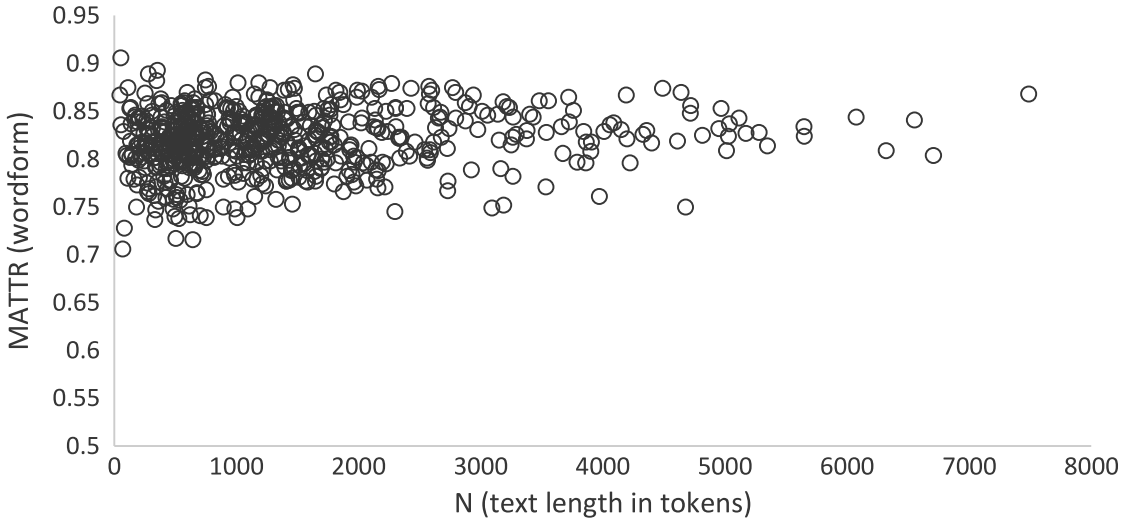

Fig. 4.1 Relationship between *MATTR* (word form) and text length in Czech texts

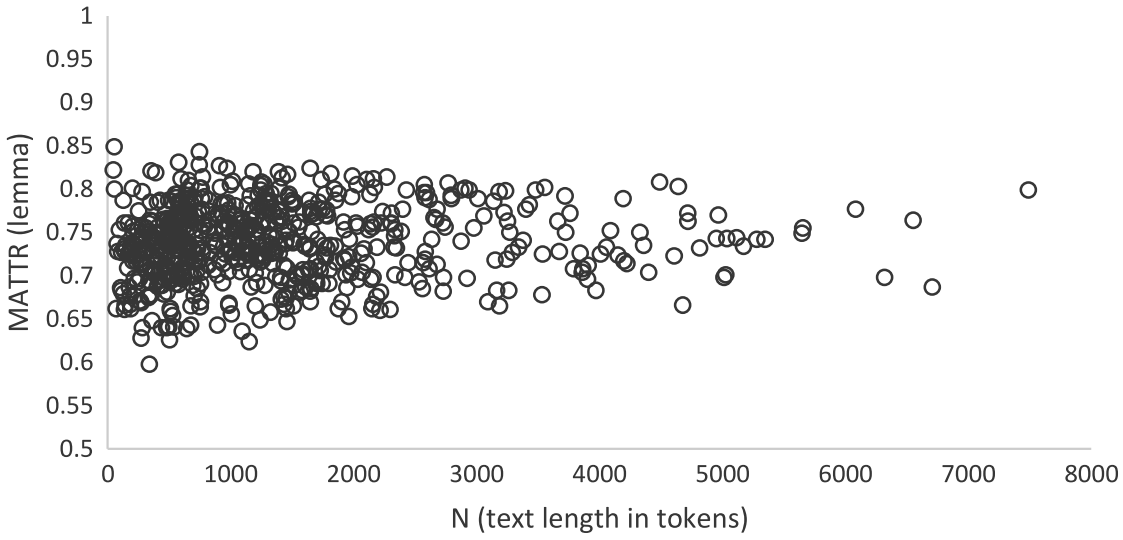

Fig. 4.2 Relationship between *MATTR* (lemma) and text length in Czech texts

another. Consequently, *MAMR* can be considered a suitable index in stylometry, at least due to its independence from text length.

Results

In stylometry, the usefulness of any method is determined by its effectiveness for a given text classification task (Juola, 2008; Kubát, 2016). In this study, we focus on two kinds of text classification: genre and authorship analysis. Our aim is to apply *MAMR* to presorted groups of texts and to observe whether significant differences appear between pairs of groups. If so, we can state that *MAMR* reflects a property of

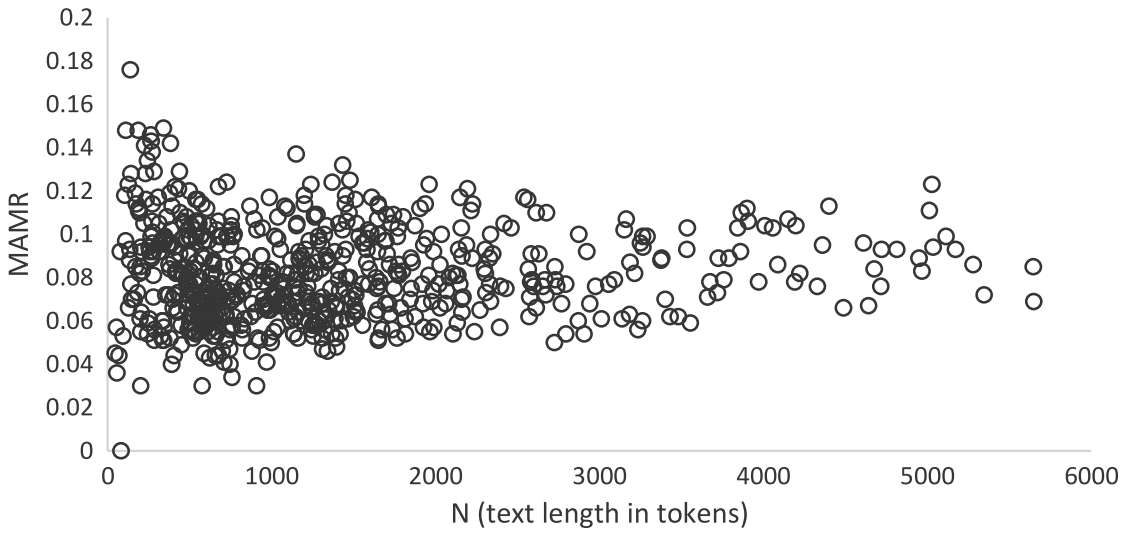

Fig. 4.3 Relationship between *MAMR* and text length in Czech texts

Table 4.1 The average *MAMR*, standard deviation (*s*), number of texts (*n*), and the adjusted *p*-values of *u*-test by genre (adjusted by the Benjamini–Hochberg–Yekutieli procedure)

		Travel books	Letters	Short stories	Novels	Newspaper articles
MAMR		0.066	0.103	0.098	0.078	0.080
s		0.017	0.020	0.016	0.016	0.022
n		132	93	71	80	92
U-test results by genre	Letters	**<0.001**				
	Short stories	**<0.001**	0.364			
	Novels	**<0.001**	**<0.001**	**<0.001**		
	Newspaper articles	**<0.001**	**<0.001**	**<0.001**	>0.999	

Bolded values denote a significant difference ($\alpha < 0.05$)

text group(s), which are strongly influenced by pragmatic factors, such as genre or authorship. In this study, the window size is set at $L = 100$.[8]

Genres

There are five genres (travel book, letter, short story, novel, and newspaper article) used for analysis in the current study. For each text, the *MAMR* is computed, and then the mean of the *MAMR* for the particular genre is determined. The results are presented in Table 4.1 and Fig. 4.4. The differences are obvious at first sight. *MAMR*

[8]The value $L = 100$ is chosen arbitrarily based on its usefulness in the previous analyses of this textual property.

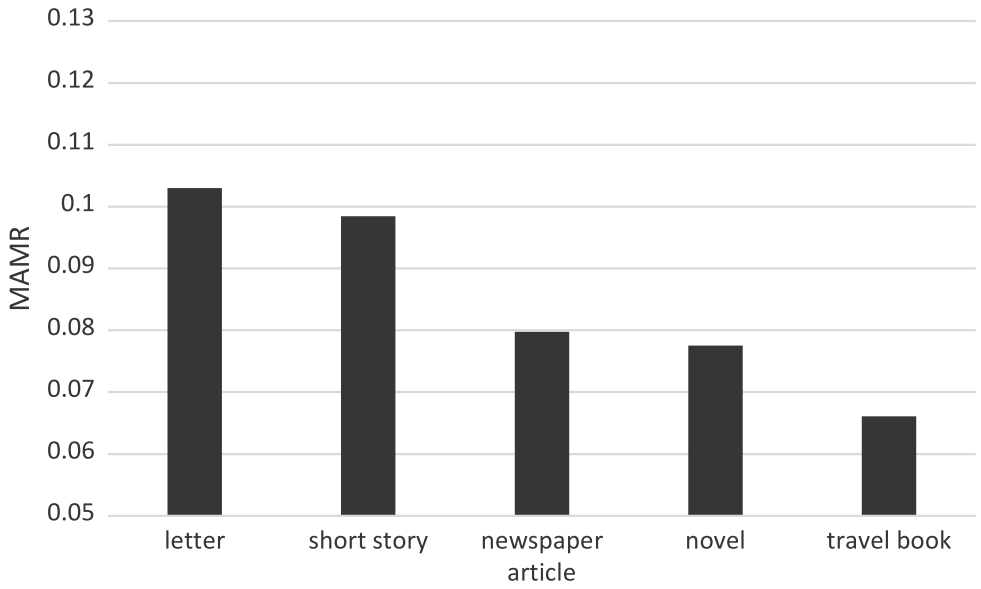

Fig. 4.4 Average *MAMR* results by genre

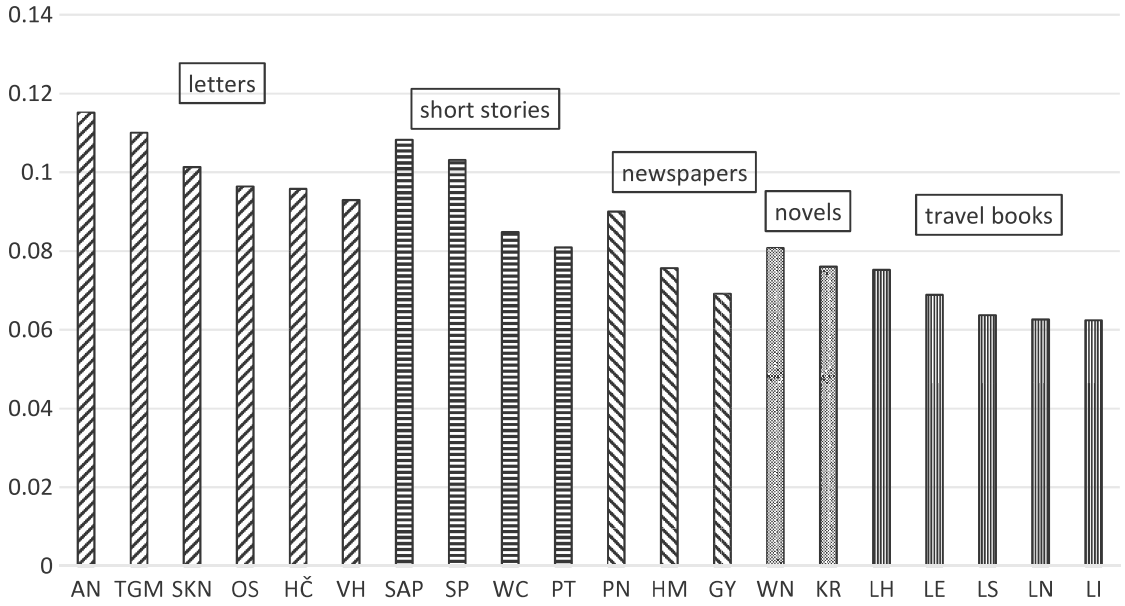

Fig. 4.5 Average *MAMR* results in Karel Čapek's books (i.e., individual novels, travel books, short story collections, etc.)

also reveals minimal differences between letters and short stories as well as between newspaper articles and novels. The observed similarities are not just "optical" (see Fig. 4.4); the results of statistical testing confirm nonsignificant differences between these pairs of groups (Table 4.1). For more details, the results of average *MAMR* for individual books are presented in Fig. 4.5.

Table 4.2 The average *MAMR*, standard deviation (*s*), number of texts (*n*), and the adjusted *p*-values of *u*-test in authorship (adjusted by the Benjamini–Hochberg–Yekutieli procedure)

	Jirásek	Němcová	Vančura	Čapek	Hrabal	Poláček	Čech
MAMR	0.078	0.103	0.065	0.078	0.087	0.074	0.067
s	0.008	0.011	0.011	0.016	0.008	0.021	0.008
n	43	19	15	80	16	74	28
Němcová	**<0.001**						
Vančura	**<0.001**	**<0.001**					
Čapek	>0.999	**<0.001**	**<0.001**				
Hrabal	**<0.001**	**<0.001**	**<0.001**	**<0.001**			
Poláček	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**		
Čech	**<0.001**	**<0.001**	**<0.001**	**<0.001**	>0.999	**<0.001**	

Bolded values denote a significant difference ($\alpha < 0.05$)

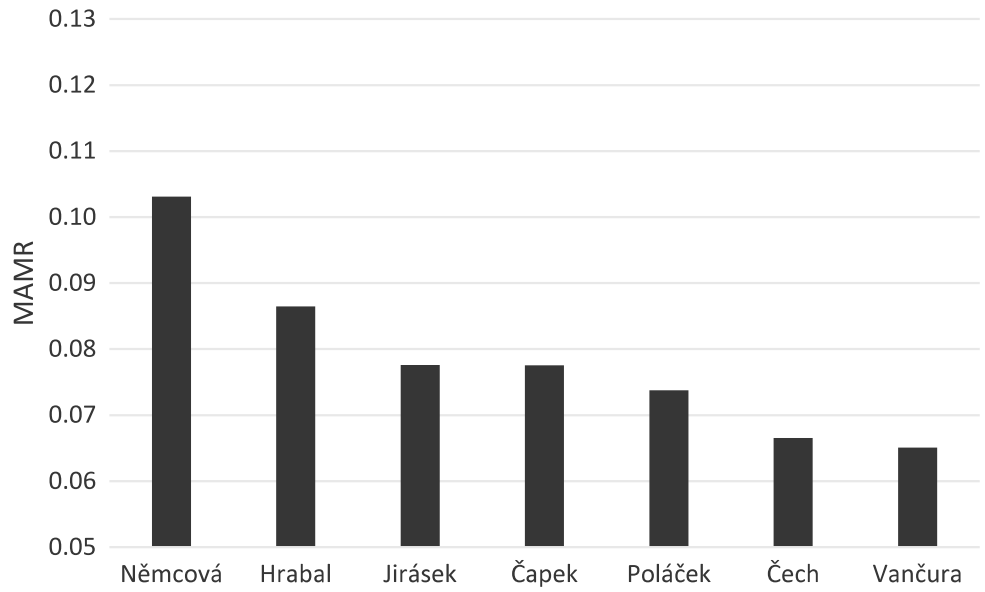

Fig. 4.6 Results of the average *MAMR* of eight Czech novelists

Authorship Analysis

For the purpose of authorship attribution, texts of one specific genre, i.e., novels, were selected. The chosen authors represent a varied spectrum of Czech writers — they were active from the middle of the nineteenth century to the second half of the twentieth century; some of them are identifiable by readers due to their specific style of writing (particularly, Vančura and Hrabal). As can be seen in Table 4.2 and Fig. 4.6, the *MAMR* results reflect significant differences between most pairs of authors.[9] Moreover, the *p*-values indicate great (and unexpected) differences among the particular authors. Consequently, MAMR seems to detect some important aspects of authorship attribution, at least among the novelists.

[9] Except two of them (Jirásek vs. Čapek and Čech vs. Hrabal).

Conclusion

Moving Average Morphological Richness is a method of measuring morphological complexity that offers intelligible interpretation and is, moreover, independent from text length. Given that the majority of the differences found by the study are significant (in genres 8/10 = 80%, in authorship 19/21 = 90.5%), the proposed method can be considered a promising stylometric tool (especially, for the analysis of a group of texts). In genre classification of Čapek's texts, *MAMR* is more effective than the *MATTR*, thematic concentration, activity of text, and other stylometric features (cf. Kubát, 2016). Most importantly, *MAMR's* independence from text size allows us to compare texts of different lengths.

The proposed method offers the potential to uncover some unexpected stylistic properties. These findings can inspire scholars not only in linguistics (both in quantitative and qualitative stylistics) but also in literary criticism. The next step is to conduct a deeper investigation of the differences between genres and authors involving specialists in literary studies. It should be emphasized that collaboration between quantitative and qualitative researchers is necessary in this field. Quantitative stylometry only provides some findings that should be subsequently interpreted from a qualitative point of view; otherwise, the obtained results can only be used for automatic text classification. This work is the first attempt to discuss whether *MAMR* is a suitable feature for stylometric research. Therefore, stylometric research using *MAMR* is a matter for further study.

Appendix

List of Texts Used for the Genre Analysis

Author	Genre	English title	Czech title	Tag
Karel Čapek	Travel book	*Letters from England*	*Anglické listy*	LE
		Letters from North	*Cesta na sever*	LN
		Letters from Italy	*Italské listy*	LI
		Letters from Holland	*Obrázky z Holandska*	LH
		Letters from Spain	*Výlet do Španěl*	LS
	Letter	*to Anna Nešporová*	*Anna Nešporová*	AN
		to Helena Čapková	*Helena Čapková*	HČ
		to Stanislav Kostka Neumann	*Stanislav Kostka Neumann*	SKN
		to Olga Scheinpflugová	*Olga Scheinpflugová*	OS
		to Tomáš Garrigue Masaryk	*Tomáš Garrigue Masaryk*	TGM
		to Věra Hrůzová	*Věra Hrůzová*	VH
	Short story	*Wayside Crosses*	*Boží Muka*	WC
		Stories from a Pocket	*Povídky z jedné kapsy*	SP
		Stories from Another Pocket	*Povídky z druhé kapsy*	SAP
		Painful tales	*Trapné povídky*	PT
	Novel	*Krakatit*	*Krakatit*	KR
		War with the Newts	*Válka s mloky*	WN
	Newspaper article	*How it is Made*	*Jak se co dělá*	HM
		Selected articles from The People's Newspaper	*Vybrané články z Lidových novin*	PN
		The Gardener's Year	*Zahradníkův rok*	GY

List of Texts Used for the Authorship Analysis

Author	English title	Czech title	Tag
Alois Jirásek	*Gaudeamus igitur*	*Filosofská historie*	GI
	Dog's Heads	*Psohlavci*	DH
Božena Němcová	*The Grandmother*	*Babička*	GM
	The village under mountains	*Pohorská vesnice*	VM
Vladislav Vančura	*Baker Jan Marhoul*	*Pekař Jan Marhoul*	BJM
	Last Judgement	*Poslední soud*	LJ

Author	English title	Czech title	Tag
Bohumil Hrabal	*I Served the King of England*	*Obsluhoval jsem anglického krále*	KE
	Cutting It Short	*Postřižiny*	CIS
Karel Poláček	*A House in the Suburbs*	*Dům na předměstí*	HS
	County Town	*Okresní město*	CT
Svatopluk Čech	*The Excursions of Mr. Brouček to the 15th Century*	*Nový epochální výlet pana Broučka, tentokráte do XV. století*	EC
	The Excursions of Mr. Brouček to the Moon	*Pravý výlet pana Broučka do Měsíce*	EM
Karel Čapek	*Krakatit*	*Krakatit*	KR
	War with the Newts	*Válka s mloky*	WN

References

Baerman, M., Brown, D., & Corbett, G. (Eds.). (2015). *Understanding and measuring morphological complexity*. New York: Oxford University Press.

Bane, M. (2008). Quantifying and measuring morphological complexity. In C. B. Chang & H. J. Haynie (Eds.), *Proceedings of the 26th West Coast Conference on formal linguistics* (pp. 69–76). Somerville, MA: Cascadilla Proceedings Project.

Bentz C., Ruzsics, T., Koplenig, A., Samardžić, T. (2016). Comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC) at the 26th International Conference on Computational Linguistics (COLING 2016)*, Osaka.

Březina, V., Pallotti, G. (2016). Morphological complexity in written L2 texts. *Second language research,* DOI: https://doi.org/10.1177/0267658316643125.

Čech, R. (2016). *Tematická koncentrace textu v češtině* [*Thematic concentration of text in Czech*]. Praha, Czech Republic: ÚFAL.

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.

Cvrček, V., & Chlumská, L. (2015). Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics, 39*(3), 309–325.

Cvrček, V., & Václavík, J. (2015). Jednoznačnost a kontext. Kvantitativní studie [*Unambiguity and context. A quantitative study*]. *Korpus—gramatika—axiologie, 11*(2015), 28–41.

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

Indrisano, R., & Squire, J. R. (Eds.). (2000). *Perspectives on writing: Research, theory, and practice*. Newark, NJ: International Reading Association.

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval, 1*(3), 233–334.

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics, 21*(3), 223–245.

Kubát, M. (2016). *Kvantitativní analýza žánrů* [*Quantitative analysis of genres*]. Ostrava, Czech Republic: Ostravská univerzita.

Kubát, M., Matlach, V., & Čech, R. (2014). *QUITA–Quantitative index text analyzer*. Lüdensheid, Germany: RAM.

Kubát, M., & Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics, 20*(4), 339–349.

Pinker, S. (2010). *The language instinct: How the mind creates language*. New York: Harper Collins.
Popescu, I. I., & Altmann, G. (2007). Writer's view of text generation. *Glottometrics, 15*, 71–81.
Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., et al. (2009). *Word frequency studies*. Berlin, Germany: Mouton de Gruyter.
Popescu, I. I., Čech, R., & Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid, Germany: RAM.
Scott, M. (2013). *WordSmith tools*. Liverpool, UK: Lexical Analysis Software.
Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., et al. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language, 31*(4), 461–479.
Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge: The University Press.