Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches

Miroslav Kubát
Department of Czech Language, University of Ostrava, Czech Republic

Ján Mačutek
Department of Applied Mathematics and Statistics, Comenius University, Slovakia and Department of Mathematics, Constantine the Philosopher University, Slovakia

Radek Čech
Department of Czech Language, University of Ostrava, Czech Republic

Abstract

Annual speeches of Czech and Czechoslovak presidents on the occasion of the end of the year are analyzed in this study. Several stylometric methods are used, namely, vocabulary richness expressed by the moving-average type–token ratio, an index of text activity, mean word length, mean verb distance, and cluster analysis of the most frequent words. The authors primarily focus on the differences between presidential addresses given in the democratic and in the communist era. Peculiarities of individual styles of particular presidents are also investigated.

Correspondence:
Miroslav Kubát, Department of Czech Language, University of Ostrava, Reální 5, Ostrava, 701 03, Czech Republic.
E-mail:
miroslav.kubat@gmail.com

1 Introduction

The aim of this study is to investigate annual speeches of Czech and Czechoslovak presidents delivered on the occasion of the end of the year from the viewpoint of several stylometric indices (vocabulary richness, text activity, mean word length (MWL), mean verb distance (MVD), cluster analysis of the most frequent words (MFWs). These methods were chosen because—unlike many other text indices—they do not depend on text length and because of their effectiveness in stylometric research (Kubát *et al.*, 2014; Kubát, 2016). Especially indicators based on type–token ratios, such as lexical richness, can be sensitive to text sizes because longer texts tend to repeat words and thus to keep the vocabulary size limited (Kubát *et al.*, 2014). Since our corpus consists of texts of various lengths (Fig. 1), we selected indicators that are less sensitive to text size (this issue is discussed in detail in Section 4). The main goal is to discover whether addresses given by presidents in the democratic and in the communistic era differ not only in their contents and vocabulary (which is obvious) but also with respect to their styles. As an interesting 'by-product' we obtained also some characteristics of the individual style of each president.

This study follows several analyses (not necessarily making use of quantitative methods) of political speeches in the USA (Lim, 2004; Savoy, 2010, 2015a,b,c, 2016, 2017, 2018a,b; Liu, 2012; Matić,

Digital Scholarship in the Humanities © The Author(s) 2020. Published by Oxford University Press on behalf of EADH.

doi:10.1093/llc/fqz089

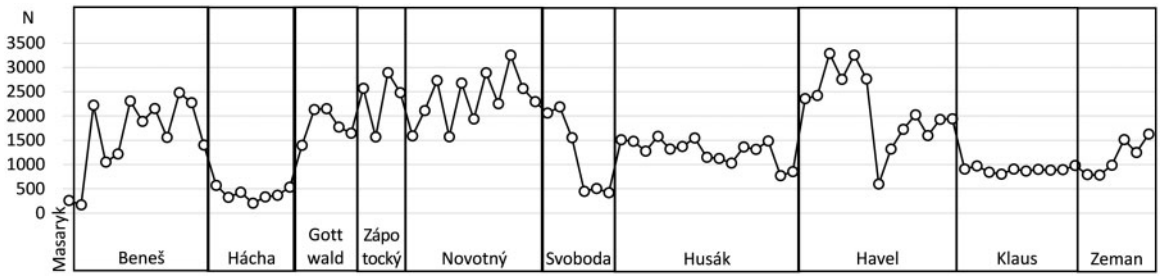

Fig. 1 Chronologically ordered text lengths (*N*) of presidential annual addresses

2012; Kubát and Čech, 2016; Quam and Ryshina-Pankova, 2016; Garcia, 2018; Wang and Liu, 2018; the paper by Chen and Liu (2015) is also closely related), Czechoslovakia and Czech Republic (David *et al.*, 2013; Čech, 2014; Fidler and Cvrček, 2015; Jičínský and Marek, 2017), Italy (Tuzzi *et al.*, 2010), France (Van Noije and Hijmans, 2005; Choi and Benoit, 2013; Roitman 2014), and Russia (Kuznetsova, 2016). Within the context of the Czechoslovak and Czech history, promising results in terms of distinguishing communist and democratic era presidents with respect to the style of presidential speeches were obtained especially by David *et al.* (2013) and by Čech (2014). In these works, the annual presidential speeches from years 1949 to 2011 are analyzed with the emphasis given on thematic concentration and the so-called thematic words (Popescu *et al.*, 2009), while we focus on the style, (almost) independently from the speech topic, contents, etc.

This article is organized as follows. The review of literature relevant to this topic is provided in Section 1. In Section 2, the history of Czechoslovakia and of the Czech Republic is briefly sketched, and the list of Czechoslovak and Czech presidents is presented. Section 3 describes the corpus of the eighty-nine presidential speeches that are analyzed in this article. In Section 4, the stylometric indices chosen by the authors are introduced and applied to the speeches. Finally, Section 5 summarizes the results and interprets them in light of the historical background.

2 Historical Background

The independent country of Czechoslovakia was established in 1918, after the collapse of the Habsburg empire at the end of World War I. According to its constitution valid in 1918–38, the Czechoslovak Republic was a parliamentary democracy (we note that in addition to today's territories of the Czech Republic and the Slovak Republic, interwar Czechoslovakia included also the region of the Carpathian Ruthenia, which is nowadays a part of Ukraine). The country enjoyed a reputation of being politically and economically stable in those years (at least relatively, if compared with much more turbulent history of its neighbors). However, it has its share of problems, as the political and economic power was concentrated in the western, i.e. Czech part of the state; in addition, there were also (not totally unrelated) ethnic tensions, and Slovaks and minorities of Germans, Hungarians, Poles, and Ruthenians struggled for more autonomy.

One of the decisive points in the history of Czechoslovakia was the year 1938 when Nazi Germany incorporated the Sudetenland (the areas along the southern, western, and northern border of today's Czech Republic). This region was a part of Czechoslovakia, but it was inhabited mostly by ethnic Germans. In the next year, Slovakia declared its independence (at the cost of having to cede some regions to Hungary) and the rest of Czechoslovakia (today's Czech Republic without the Sudetenland) became a German-dependent territory known as the Protectorate of Bohemia and Moravia. The Protectorate had its own state president (the office was held by Emil Hácha), although the Reich Protector appointed by Berlin had the highest authority.

After World War II, Czechoslovakia fell within the Soviet sphere of influence (and had to cede the Carpathian Ruthenia to the Soviet Union). The Communist Party failed to achieve a decisive

victory in the 1946 election, but Soviet-backed communists then took power in a coup d'état in 1948. Czechoslovakia was declared a people's democracy and later, in 1960, it changed its official name to the Czechoslovak Socialist Republic. The ideological principles of Marxism-Leninism formed a basis in economy (central planning, with private ownership being effectively abolished), politics (according to the 1960 constitution, the Communist Party was the leading political force) as well as in everyday life (especially religious and cultural activities that were not in line with the official socialist/communist ideology were suspicious, if not directly forbidden). The communist regime was brutal especially in the 1950s, when a number of people who were considered as a threat to the official ideology were imprisoned or executed.

The regime slightly softened in the mid-1960s, and especially in 1968 (the period known as the 'Prague Spring'), when relatively liberal people were elected to key posts in the Communist Party. The constitution was substantially amended, and the unitary Czechoslovak state became a federation of the Czech Socialist Republic and the Slovak Socialist Republic. The call for reforms could be heard (minor one actually took place, such as loosening restrictions on media and speech) and the overall atmosphere changed. However, concerns about such a development (and about a threat of losing the influence) arose in the Soviet Union. Communist hard-liners in Czechoslovakia also opposed the changes in society. The complicated situation escalated in August 1968, when negotiations failed to reach a compromise that would be acceptable for both sides. Armies of several Warsaw Pact countries invaded Czechoslovakia, ending thus any further liberalization attempts.

The time after the invasion is known as the 'normalization', and especially years closely following 1968 present another peak in political repressions (albeit weaker than in the 1950s). Many Czechoslovak citizens (above all educated ones) emigrated to the West. However, communists did not achieve complete control of internal affairs. In 1970s and 1980s, political (especially in today's Czech Republic) and religious (especially in today's Slovakia) activists emerged as opposition that, although without any political power, was able to challenge the regime. A period of an economic stagnation also contributed to the weakening of the rule of the Communist party.

The communist regime in Czechoslovakia finally fell in 1989. After Gorbachev's rise to power, the Soviet Union was no longer willing (and maybe able) to support communists in its satellite states (Czechoslovakia being one of them). The Velvet Revolution in autumn 1989 resulted in a non-violent transition of power and in many far-reaching changes. In 1990, the first democratic elections after 1946 were held.

In the new circumstances, Slovak demands for a greater autonomy gradually led to a non-functioning federation. Czechoslovakia was peacefully dissolved at the end of 1992, and two newly independent countries, the Czech Republic and the Slovak Republic, were founded on 1 January 1993. Both new countries inherited—and, with some minor changes, still keep—the political system of democratic Czechoslovakia. They are parliamentary representative democracies. The head of the government is the prime minister, while the president is a formal head of state (with limited and specific powers).

In the years 1949–89, the political monopoly was held by the Communist Party of Czechoslovakia (KSČ) that was led by the general secretary. The general secretary was the one holding the most powerful office in the state and the role of the president was rather representative. It is important to mention that both offices, the general secretary and president, were sometimes served by the same person (Klement Gottwald, Antonín Novotný, Gustáv Husák). Given that Czechoslovakia in 1949–89 was a satellite state of the Soviet Union, the real power of all positions was limited to some extent. Important decisions were highly influenced by the Soviet Union.

The role of the president according to the Constitution of the Czech Republic has been rather representative since 1989, however, he/she still has a considerable role in political affairs. The prime minister is the most important leader who acts as the head of the government. Given that many powers have to be signed by both the president and the prime minister of the Czech Republic, responsibility for some political issues is effectively shared between the two offices. The role of the president has considerably increased since 2013 due to the change of the election procedure. Until 2012, the president was elected by the parliament. Since 2013 the president is elected by direct popular vote. Although the rights given by the constitution have not been changed, the real power of the

Table 1. The list of Czechoslovak and Czech presidents

Name	In-office
Tomáš Garrigue Masaryk	1918–35
Edvard Beneš	1935–38, 1940–48
Emil Hácha	1939–45
Klement Gottwald	1948–53
Antonín Zápotocký	1953–57
Antonín Novotný	1957–68
Ludvík Svoboda	1968–75
Gustáv Husák	1975–89
Václav Havel	1989–2003
Václav Klaus	2003–13
Miloš Zeman	2013ff

president has increased. Miloš Zeman, the first president elected directly by people, often emphasizes this fact as an argument to strengthen his position.

To sum up, Czechoslovakia (and the Czech Republic as one of its two succession states) was undisputedly a democratic republic in 1918–38 and since 1989. We can thus distinguish three groups of Czechoslovak and Czech presidents. There are five Democratic presidents (Tomáš Garrigue Masaryk and Edvard Beneš in the interwar Czechoslovakia, Václav Havel, Václav Klaus, and Miloš Zeman after the fall of the communist regime in 1989) and five communist presidents (Klement Gottwald, Antonín Zápotocký, Antonín Novotný, Ludvík Svoboda, and Gustáv Husák) in the history of Czechoslovakia and of the Czech Republic. Emil Hácha has quite an extraordinary position because, during his office (1939–45), the country (under the name of the Protectorate of Bohemia and Moravia) was under the control of Nazi Germany. Therefore, speeches by Hácha are considered as a special group in our corpus. The list of Czech and Czechoslovak presidents can be seen in Table 1. It must be noted that President Beneš resigned from his office under German pressure in 1938. However, he later acted as the head of the state in exile, and as such he delivered also several presidential speeches. We thus have 5 years (1941–45) from which we analyze two 'parallel' presidential speeches, by Beneš and by Hácha (see the Appendix). Another exception is year 1993 when Czechoslovakia was dissolved, and two newly independent countries were created—President Havel abdicated from his office of the Czechoslovak head of state in July 1992 and was elected the president of the Czech Republic in the second half of January 1993, hence the speech was not delivered in that year.

3 Language Material

The corpus under analysis consists of 89 annual speeches of all eleven Czechoslovak and Czech presidents. The addresses were delivered on the occasion of the end of the year since 1935 when the first Czechoslovak President Tomáš Garrigue Masaryk started this relatively long tradition (incidentally, Masaryk delivered only one speech, as he resigned from his office on December 1935 because of his age and health). The speeches are addressed to the Czechoslovak (later to Czech) citizens and are broadcasted by radio and television. Presidents primarily summarize important events from the past year, and secondarily look ahead toward issues of the upcoming year (although some speeches by Hácha can be considered as Christmas messages rather than political reports—he focused more on Christmas topics like family, children, God, Christianity, goodwill, fellowship, etc., probably in order to avoid or to moderate the negative themes of the wartime). It is also important to mention that except for Gustáv Husák, all presidents were/are Czech native speakers. Although Husák was Slovak and his mother tongue was Slovak, he gave his annual presidential speeches sometimes in Czech and sometimes in Slovak. These two languages are similar, and all the transcriptions are provided in Czech. It is even not clear in which language the speeches were written originally. According to the results, this factor does not have any significant influence on our research.

Almost all texts were obtained from the collection of Český rozhlas, the Czech public radio broadcaster that has operated since 1923 (https://interaktivni.rozhlas.cz/prezidentske-projevy/). There are both texts and recordings of the speeches, the only exception being the address given by Emil Hácha in 1939 for which only the recording is available (its transcription was made by the authors). The last speech delivered by Zeman (2019) was obtained from the official website of the Office of the President of the Czech Republic (https://www.hrad.cz/). It is important to mention that we work with a whole statistical population because all speeches of this particular genre are included in our corpus (statistical tests, which make an

inference from a sample to the respective population, thus cannot be applied here). Considering all the specifics of the Czechoslovak and Czech presidential addresses, it can be said that this genre is very homogenous and appropriate for this stylometric analysis. Tuzzi *et al.* (2010, p. 3) who analyzed the collection of 60 end-of-year Italian presidential speeches described this type of data as follows: 'A corpus having this character is a gate to an infinite paradise for textologists.'

Lengths of individual texts (measured in the number of words) and the mean values for all presidents can be seen in Figs 1 and 2, respectively. The text lengths display a high level of heterogeneity not only in the corpus as the whole, but also within subgroups created by speeches delivered by the same president (especially by Beneš, Svoboda, and Havel, see Fig. 1). We remind that many text indices depend on text length, even if this fact does not have to be too obvious (Čech, 2015; Mitchell, 2015). Given the abovementioned heterogeneity, one must be cautious in the choice of stylometric methods.

4 Methodology and Results

The indices described in this section were calculated by Quantitative Index Text Analyzer (QUITA; Kubát *et al.*, 2014) and moving-average type–token ratio (MATTR) software (Covington and McFall, 2010). The cluster analysis of the MFWs was performed by Stylo R package (Eder *et al.*, 2016). The corpus was lemmatized, so lemma is considered as the basic unit in this study. Lemmatization was done by the program Treex that is developed by the Institute of Formal and Applied Linguistics, Charles University, Prague (Popel and Žabokrtský, 2010). To be more specific, we used the web interface of Treex available at http://ufal.mff.cuni.cz/treex.

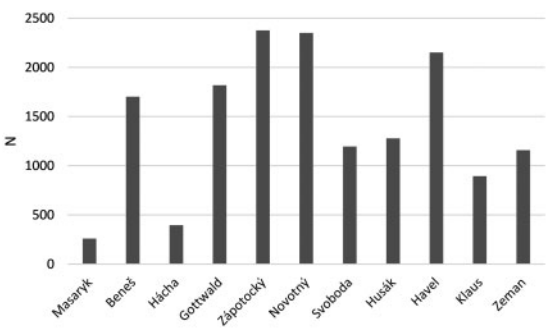

Fig. 2 Chronologically ordered mean text lengths (N) of presidential annual addresses

The list of all speeches with the corresponding numerical values of the indices can be found in the Appendix of this article.

4.1 MATTR

The measurement of vocabulary richness is one of the oldest quantitative methods in stylometry with more than 70 years long history (Popescu *et al.*, 2009). A large number of indices of vocabulary richness were set up in linguistics; however, almost all of them are dependent on the text length (e.g. Mitchell, 2015, compared fifty type–token formulas, coming to the conclusion that there is a relation between the type–token ratio and the text length). Given that the length of particular presidential speeches is far from being constant (Figs 1 and 2), we have, on the one hand, some standard measures of vocabulary richness (e.g. the type–token relation, see Wimmer, 2005; Mitchell, 2015), but, on the other hand, most of them would reflect the text length rather than—in our case—the style of the president who delivered the speech. To avoid this dependence in our analysis, we decided to use the MATTR proposed by Covington and McFall (2010), which was experimentally proved to be independent of the text size (Kubát and Milička, 2013; Kub-t, 2014, 2016). We note that in addition to vocabulary richness there is also another accepted interpretation of the type–token relation—it can be understood as a model of the information flow in a text (Wimmer, 2005, p. 361).

The MATTR is defined as follows. A text is divided into overlapping subtexts of the same length (the so-called 'windows' with an arbitrarily chosen size L). The window moves forward one token at a time. The type–token ratio is computed for every window and, finally, the MATTR is defined as a mean of the type–token window values, i.e.

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

where L is the length of the window, $L < N$, with N being the text length in tokens, and V_i is the number of types in the ith window.

Let us demonstrate the calculation of MATTR on a short text:

John loves Mary. Mary and John love beer.

After lemmatization, the text is as follows:

john love mary. mary and john love beer.

The text consists of eight tokens ($N = 8$) and five types ($V = 5$). Since we arbitrarily decide the window size as four tokens ($L = 4$), we get five following windows:

(1) john love mary. mary ($V = 3$)
(2) love mary. mary and ($V = 3$)
(3) mary. mary and john ($V = 3$)
(4) *mary and john love* ($V = 4$)
(5) *and John love beer* ($V = 4$)

The calculation is as follows:

$$\mathrm{MATTR}(4) = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)} = \frac{3+3+3+4+4}{4(8-4+1)} = 0.85$$

In our analysis, we work with the window of the size $L = 100$ words. This value of L was used by Kubát and Milička (2013) and by Kub-t (2014, 2016). While there are no objective rules for the choice of the window size, Covington and McFall (2010, pp. 97–99) dedicate a section of their paper to this question. They prefer the size of $L = 500$ words, and their recommendation is that it should be 'smaller than the smallest text to be processed, but large enough to provide a meaningful measure of style'. Short windows are appropriate for other research goals, such as the detection of repetitions of words or phrases. Given that the shortest text in our corpus consists of 107 words (the speech by Beneš in 1936, see the Appendix), the size of $L = 100$ words seems reasonable.

Figures 3 and 4 present the values of the MATTR for all speeches under analysis and the mean values for particular presidents, respectively.

4.2 Text activity (Q)

A text can give a stronger emphasis either on the action (plot) or on the description. For instance, travel books focus principally on description, while short stories usually concentrate on the plot. The concept of the activity and descriptivity was introduced by Busemann (1925). It is assumed that the text activity is represented by verbs and the descriptivity by adjectives. The index of activity Q is defined as a ratio of the number of verbs to the sum of the number of verbs and adjectives in the text, i.e.

$$Q = \frac{V}{V+A},$$

with V being the number of verbs and A the number of adjectives in the text (Altmann, 1978).

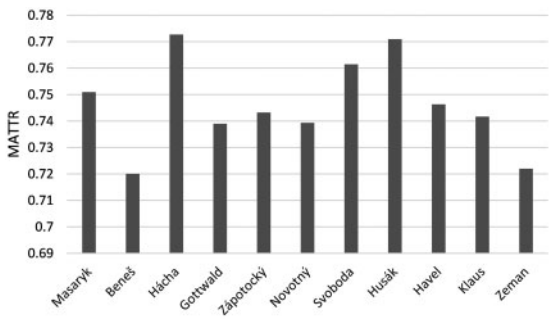

Fig. 4 Chronologically ordered mean values of vocabulary richness (measured by the MATTR) of presidential annual addresses

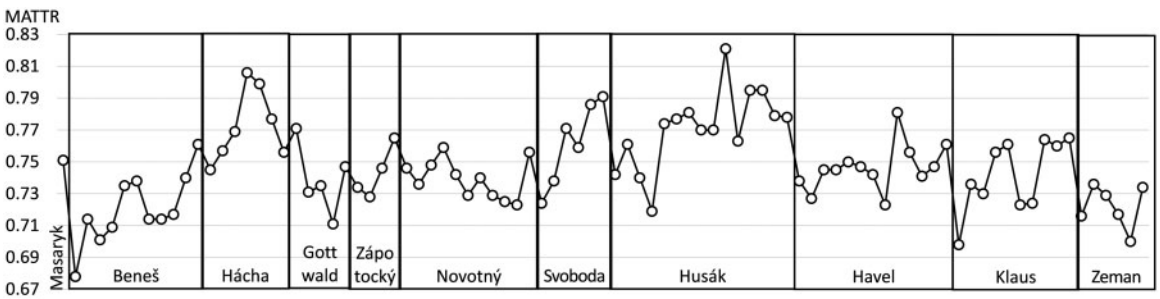

Fig. 3 Chronologically ordered values of vocabulary richness (measured by the MATTR) of presidential annual addresses

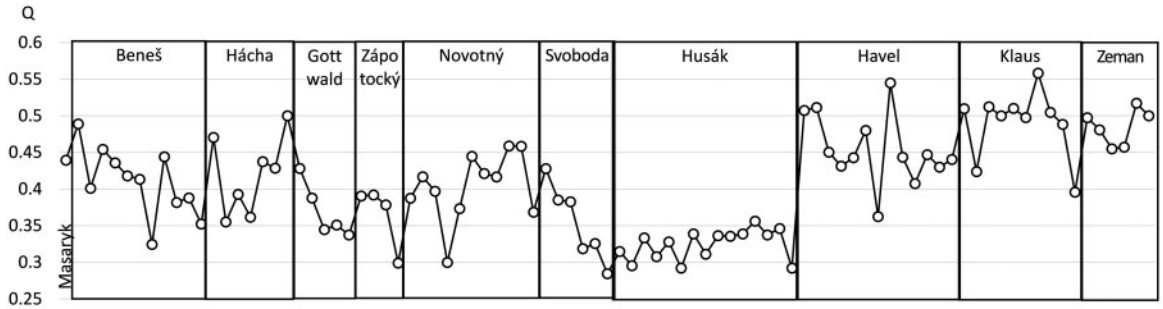

Fig. 5 Chronologically ordered values of activity (Q) of presidential annual addresses

Admittedly, the index Q as defined above is a very rough measure of the text activity (e.g. not all verbs express activity to the same extent). However, due to its very simple form, it found its use in text analysis (for relatively recent works, see e.g. Těšitelová, 1992, p. 112; Popescu *et al.*, 2013, 2014b, 2015, pp. 216–31; Altmann and Köhler, 2015, pp. 17–31; Zörnig *et al.*, 2015, p. 4). Zörnig *et al.* (2015) provide references to older works in which this index was applied. Těšitelová (1992, pp. 110–15) mentions several similar text indices.

Let us demonstrate the calculation of Q on a short text:

John is sad. Mary is happy and lucky.

There are two verbs: is, is, ($V = 2$), and three adjectives: sad, happy, lucky, ($A = 3$) in the text.

The calculation is as follows:

$$Q = \frac{V}{V + A} = \frac{2}{2 + 3} = 0.4$$

As can be seen in Figs 5 and 6, low values of the text activity index Q are typical for the presidents from the communist era (Gottwald, Zápotocký, Novotný, Svoboda, Husák). This observation will be discussed more in detail in Section 5.

4.3 MWL

The arithmetic mean of the lengths of words (i.e. tokens) is measured in the number of graphemes in this study. Length belongs to one of the most often-studied properties of words, but it is mostly measured in the number of syllables of which the word consists (Popescu *et al.* 2014a, pp. 14–86 and references

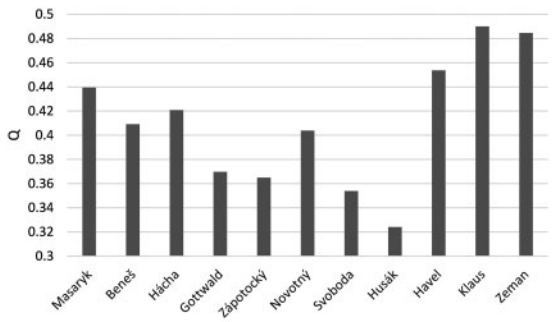

Fig. 6 Chronologically ordered mean values of activity (Q) of presidential annual addresses

therein). A definition of word length as the number of graphemes that the word contains, although rare, can be found in several works (Herdan, 1958, 1966; Alexeev, 1998; Eeg-Olofsson, 2009). It could be argued that, as we analyze speeches, it would be more appropriate to measure word length in the number of phonemes, but in Czech there are only minor differences between the number of phonemes and graphemes in words (according to Short, 1993, p. 459, '[t]he relationship between phonemes and alphabet is close'). In our article, the MWL is defined as

$$\mathrm{MWL} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

where N is the number of tokens in the text and x_i is the length of the ith token in graphemes.

The observed values of the MWL (Figs 7 and 8) again distinguish the communist presidents from the democratic era. The former ones tend to use longer

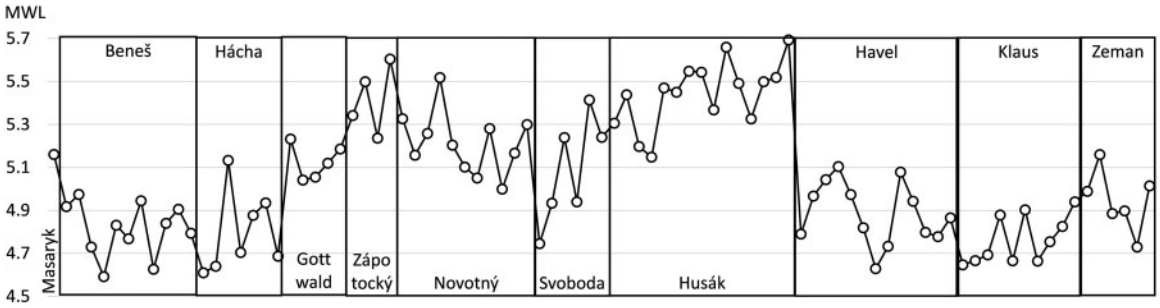

Fig. 7 Chronologically ordered values of MWL of presidential annual addresses

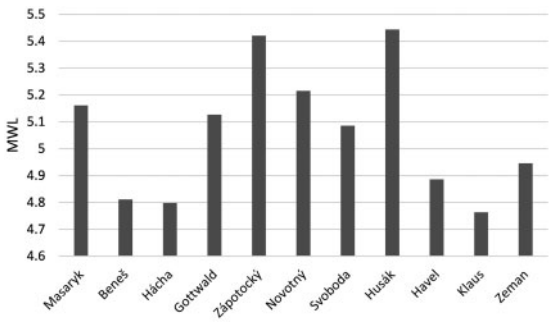

Fig. 8 Chronologically ordered mean values of MWL of presidential annual addresses

words (Masaryk is an exception in this respect, but he delivered only one speech).

4.4 MVD

The MVD is a stylometric indicator that expresses how many tokens occur on average between two successive verbs in a text. It is included in the QUITA software (Kubát *et al.*, 2014). The idea behind the MVD is to find an indicator that is both easy to evaluate and that at least roughly corresponds to the mean clause length (MCL). The MCL is used mainly as a measure of syntactic complexity in fields like psycholinguistics (it rises with increasing age and the level of education, see e.g. Haswell, 2000; Berman and Ravid, 2009) or the second language acquisition research (it rises with the increasing proficiency level, see e.g. Lintunen and Mäkilä, 2014; Neary-Sundquist, 2017).

There are, however, two general problems related to calculating the MCL. First, it is quite difficult to distinguish individual clauses automatically because of multifunctional usage of punctuation. Secondly, there are problems connected with finding borders between clauses in oral speeches during transcription. Given that the primary function of a verb is being the predicate in a clause, the MVD seems to be quite a reasonable tool that can be a substitute for the MCL. Needless to say, this method has its own problems, as there is a risk of an oversimplification (e.g. all verbs are taken into account without distinguishing their function in a sentence). Nevertheless, the MVD is an easy-to-obtain indicator that—as it correlates with the MCL—can serve as a characteristic of syntactic properties of a text. The usefulness and efficiency in stylometric research have been proved in genre analysis by Kubát (2016).

The speeches by the democratic presidents (especially after 1989, i.e. by Havel, Klaus, and Zeman) tend to display lower values of the MVD than those by the communist ones (Figs 9 and 10). This fact is related to the higher frequency of verbs and therefore higher activity in the democratic speeches (Section 4.2). Generally, we can say that the democratic style in the period after the year 1989 is more informal and therefore syntactically simpler.

4.5 MWFs cluster analysis

Cluster analysis in this study is based on relative frequencies of the 100 MFWs in the corpus. The so-called 'culling' was set to 60% which means that only words that appear in at least 60% of the texts in the corpus will be considered in the analysis (Eder *et al.*, 2016). As the corpus under analysis consists of forty-one speeches delivered by democratic presidents (46% of texts in our corpus), forty-one speeches by communist presidents (46%), and seven speeches by Hácha (8%), the culling ensures that the clusters are not influenced

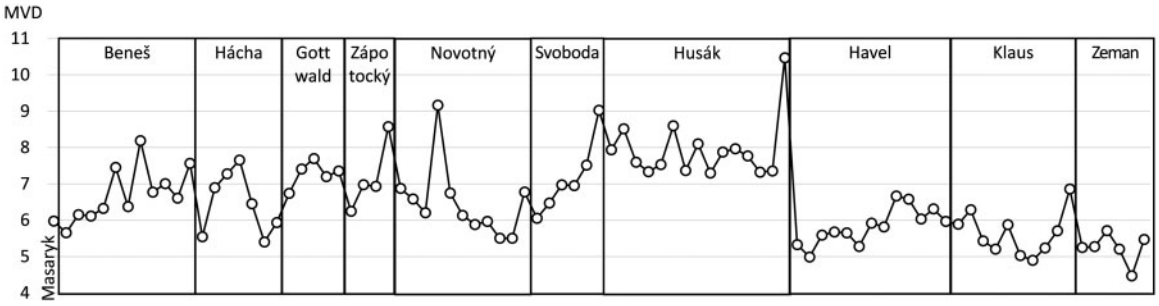

Fig. 9 Chronologically ordered values of MVD of presidential annual addresses

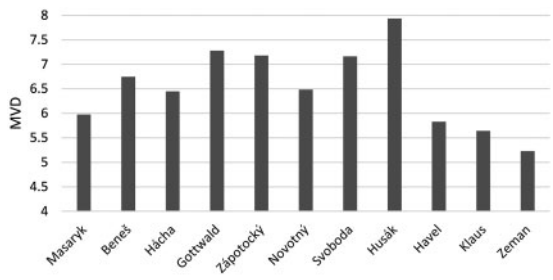

Fig. 10 Chronologically ordered mean values of MVD of presidential annual addresses

by a specific vocabulary regardless of whether Hácha is considered as a special group or whether he is attached to any of the democratic or communist presidents (presumably, especially all communist presidents repeated many words, such as communist, socialist, Soviet Union, etc.). Thus, our clustering of the speeches is based on words that are common in both (or in all three) groups.

The classic Delta distances were selected to measure similarities between frequency patterns of individual texts in the corpus. The formula for the Delta distance suggested by Burrows (2002) is

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right|,$$

where n is the number of MFWs; A, B are the compared texts; A_i and B_i are the frequencies of a given word i in the texts A and B, respectively; μ_i is the mean frequency of a given word in the corpus, and σ_i is the standard deviation of frequencies of a given word (see also Eder *et al.*, 2016).

Clusters based on the MFW are presented in Fig. 11. It can be seen that speeches by the democratic and communist presidents are placed in two parts of the dendrogram. These two groups do not overlap. The individual speeches of each president seem to be quite homogeneous with respect to the MFW and are clustered close to each other. The only exception is Ludvík Svoboda. We can see in Fig. 11 that his six speeches are divided into two groups. The first group (1969, 1970, 1971) is in one cluster with Gottwald, Zápotocký, and Novotný. The second one (1972, 1973, 1974) is close to Husák. This change in the style of the Svoboda's speeches is most likely caused by the fact that he suffered a stroke in June 1972. His health then worsened, and he left his office in 1975. That is why his speech in 1974 was read by a radio host and the speech in 1975 was even not delivered by Svoboda but by Husák (the general secretary of the Communist Party at that time) who become a president in May 1975. Svoboda's three speeches in 1972–74 are therefore much shorter than the previous ones (text lengths in tokens in years 1969–74: 2,059, 2,185, 1,551, 448, 507, 420). The content of his last three speeches is rather formal and there are no specific topics (the speeches mostly consist of a greeting and wishes of happiness, success, etc., in the year that has just begun). We assume that Svoboda hardly participated in writing of his last speeches, he probably just read the given texts.

5 Conclusion and Discussion

The results obtained (see Table 2 and the Appendix) reveal several findings that could be summarized as follows. With the exception of the MATTR (see Section 4.1), all methods used seem to be effective in terms of distinguishing a typical presidential speech style of the

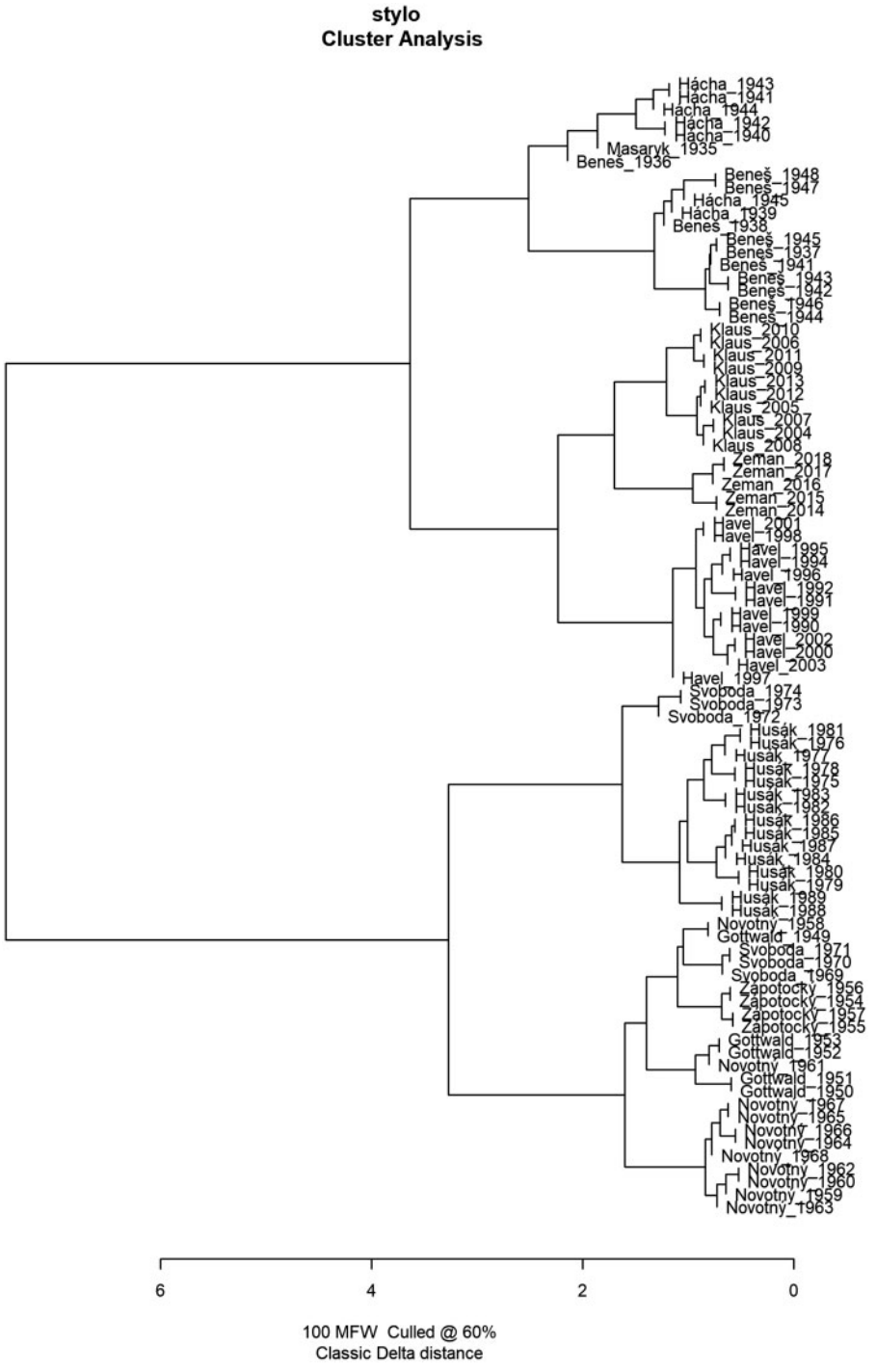

Fig. 11 MFW cluster analysis of presidential annual addresses

Table 2. Mean values of stylometric indices from annual addresses given by democratic presidents, communistic presidents, and Emil Hácha

Index	Democracy	Communism	Hácha
N	1,532	1,727	395
MATTR	0.73	0.75	0.77
Q	0.45	0.36	0.42
MWL	4.85	5.29	4.80
MVD	5.95	7.28	6.45

democratic and communist era. The MATTR index is probably a text characteristic that is independent of the political ideology, or it can be a characteristic of the genre rather than of a writer/speaker. Given the peculiarities of the Czechoslovak and modern Czech history, it is interesting to note that the style of Hácha is very close to the democratic presidents in all aspects under study except for the text length.

In particular, speeches of the communist presidents are more descriptive (less active) that is probably caused by the fact that these texts have several themes that had to be covered because of the official communist state ideology (e.g. the political situation in the world and in Czechoslovakia, friendly relations with the USSR and other communist countries, achievements in agriculture and industry, etc.). These issues are described in a very detailed way, e.g. how many tons of potatoes were harvested, how many loaves of bread were baked or how many tons of steel were produced. We assume that the annual presidential speeches in the communist era were prepared in a more official way, using a kind of a 'template' with a beforehand made structure. These texts were probably written by several people and subsequently revised so that they be ideologically 'clean'. On the other hand, the style of addresses given by democratic presidents seems to be more informal, there are not so many obligatory themes, language is simpler, and we can, therefore, expect that these texts were written by individual presidents without corrections of other people (or at least with fewer corrections). Therefore, the speeches from the democratic era are more active, they have smaller vocabulary richness, lesser MWL and shorter MVDs. The higher MWL of the communist presidents probably can be at least partially explained by their 'obligatory' vocabulary—they had to use words like, e.g. *socialistický* (socialist), *komunistický* (communist), etc., quite often. These words, typical for a communist nomenclature, have above average lengths.

Table 3. Pearson correlation coefficient (r) between indices

Index	*Q*	MVD	MWL	Tokens	Types
MATTR	−0.46	0.48	0.46	−0.26	−0.19
Q		−0.89	−0.67	−0.02	−0.01
MVD			0.63	−0.08	−0.06
MWL				0.14	0.15
Tokens					0.98

The values of the used stylometric indicators and the assumption that communistic annual speeches differ from democratic ones, in general, were also supported by the MFW cluster analysis.

Since we analyzed speeches that were delivered during the period more than 90 years long, one can expect that some changes in the style could be caused by slow changes in the Czech language. In this respect, we emphasize that our corpus consists of two democratic eras (before 1948 and after 1989, i.e. they are divided by the gap of 40 years). Considering the results obtained, we can state that the style of the presidential annual speeches is influenced more by the political stances and ideologies of the presidents than by general language changes. However, the development of the Czech language plays its role, which can be seen especially in Fig. 11. Based on the MFW cluster analysis (see Section 4.5), the democratic and communist presidents are separated into two clusters. Nevertheless, all the pre-communist Czechoslovak presidents (Masaryk, Beneš, Hácha) form one group in the 'democratic cluster', with all the post-communist presidents (Havel, Klaus, Zeman) being placed into the other group.

As some indices are supposed to correlate, we performed a basic analysis using the Pearson correlation coefficient *r*. Its values can be found in Table 3. The tight correlation between text activity and the MVD ($r = -0.89$) is not surprising—long distances between verbs leave space which is filled with other parts of speech, among them adjectives, which decrease in the index of text activity. Strong correlations can be also observed between the MWL and text activity ($r = -0.67$) and between the MWL and the MVD ($r = 0.63$). Both of them can be explained by the frequency of verbs in a text. Adjective tends to be longer than verbs, especially in speeches by the communist

presidents, hence the less frequent use of verbs (which makes a text less active) increases the MWL and decreases the MVD.

It remains an open question whether the same pattern appears in other post-communist countries (in which the tradition of the end-of-year speech exists, e.g. in Poland) in Europe that experienced an analogous historical development.

Funding

This research was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and of the Slovak Academy of Sciences (VEGA 2/0054/18 to J.M.).

Appendix

The list of all speeches with results of the indices

Text	Types	Tokens	MATTR	*Q*	MVD	MWL
1935 Masaryk	158	261	0.751	0.439	5.976	5.161
1936 Beneš	107	170	0.678	0.489	5.654	4.918
1937 Beneš	772	2,221	0.714	0.401	6.154	4.974
1938 Beneš	426	1051	0.701	0.454	6.117	4.730
1939 Hácha	295	572	0.745	0.471	5.544	4.610
1940 Hácha	189	325	0.757	0.355	6.894	4.640
1941 Beneš	499	1,218	0.709	0.436	6.323	4.592
1941 Hácha	251	429	0.769	0.393	7.276	5.133
1942 Beneš	812	2,306	0.735	0.418	7.451	4.831
1942 Hácha	152	203	0.806	0.362	7.654	4.704
1943 Beneš	696	1,887	0.738	0.413	6.372	4.768
1943 Hácha	211	335	0.799	0.438	6.451	4.878
1944 Beneš	688	2,151	0.714	0.324	8.188	4.945
1944 Hácha	224	369	0.777	0.429	5.403	4.935
1945 Beneš	572	1,559	0.714	0.444	6.773	4.627
1945 Hácha	311	534	0.756	0.500	5.933	4.687
1946 Beneš	848	2,479	0.717	0.382	7.006	4.840
1947 Beneš	850	2,272	0.74	0.388	6.606	4.905
1948 Beneš	594	1,402	0.761	0.352	7.563	4.794
1949 Gottwald	611	1,394	0.771	0.428	6.738	5.232
1950 Gottwald	808	2132	0.731	0.388	7.408	5.041
1951 Gottwald	798	2,150	0.735	0.345	7.697	5.055
1952 Gottwald	696	1,772	0.711	0.351	7.200	5.120
1953 Gottwald	663	1,645	0.747	0.337	7.355	5.186
1954 Zápotocký	887	2,569	0.734	0.391	6.244	5.342
1955 Zápotocký	599	1,566	0.728	0.392	6.978	5.499
1956 Zápotocký	1,005	2,892	0.746	0.378	6.933	5.236
1957 Zápotocký	962	2,477	0.765	0.299	8.575	5.604
1958 Novotný	610	1,590	0.746	0.388	6.873	5.327
1959 Novotný	730	2,108	0.736	0.417	6.582	5.157
1960 Novotný	868	2,726	0.748	0.397	6.205	5.259
1961 Novotný	621	1,571	0.759	0.300	9.165	5.518
1962 Novotný	914	2,675	0.742	0.373	6.749	5.204
1963 Novotný	670	1,936	0.729	0.445	6.133	5.102
1964 Novotný	893	2,889	0.74	0.421	5.882	5.051
1965 Novotný	748	2,251	0.729	0.417	5.965	5.281

(*Continued*)

(continued)

Text	Types	Tokens	MATTR	Q	MVD	MWL
1966 Novotný	963	3,250	0.725	0.459	5.503	4.999
1967 Novotný	797	2,565	0.723	0.458	5.504	5.167
1968 Novotný	780	2,293	0.756	0.369	6.775	5.300
1969 Svoboda	765	2,059	0.724	0.428	6.053	4.746
1970 Svoboda	760	2,185	0.738	0.385	6.473	4.934
1971 Svoboda	632	1,551	0.771	0.383	6.973	5.239
1972 Svoboda	236	448	0.759	0.319	6.952	4.940
1973 Svoboda	278	507	0.786	0.326	7.515	5.414
1974 Svoboda	241	420	0.791	0.284	9.023	5.240
1975 Husák	560	1,510	0.742	0.315	7.937	5.306
1976 Husák	571	1,478	0.761	0.295	8.517	5.438
1977 Husák	480	1,276	0.74	0.333	7.596	5.197
1978 Husák	557	1,581	0.719	0.308	7.332	5.148
1979 Husák	563	1,318	0.774	0.328	7.532	5.470
1980 Husák	586	1,370	0.777	0.292	8.599	5.450
1981 Husák	634	1,546	0.781	0.339	7.368	5.548
1982 Husák	492	1,150	0.77	0.311	8.099	5.543
1983 Husák	470	1,125	0.77	0.336	7.296	5.368
1984 Husák	512	1,028	0.821	0.336	7.876	5.660
1985 Husák	576	1,362	0.763	0.339	7.965	5.491
1986 Husák	567	1,312	0.795	0.356	7.768	5.326
1987 Husák	610	1,485	0.795	0.337	7.323	5.499
1988 Husák	390	771	0.779	0.346	7.356	5.519
1989 Husák	411	854	0.778	0.292	10.469	5.693
1990 Havel	880	2,355	0.738	0.507	5.326	4.791
1991 Havel	890	2,419	0.727	0.511	4.988	4.967
1992 Havel	1,160	3,284	0.745	0.450	5.590	5.044
1994 Havel	942	2,752	0.745	0.431	5.676	5.104
1995 Havel	1,094	3,252	0.75	0.443	5.653	4.974
1996 Havel	965	2,760	0.747	0.480	5.278	4.819
1997 Havel	310	598	0.742	0.362	5.920	4.629
1998 Havel	520	1,318	0.723	0.545	5.814	4.733
1999 Havel	769	1,725	0.781	0.443	6.665	5.078
2000 Havel	832	2,023	0.756	0.408	6.579	4.943
2001 Havel	657	1,595	0.741	0.447	6.034	4.798
2002 Havel	780	1,928	0.747	0.430	6.310	4.778
2003 Havel	778	1,940	0.761	0.441	5.969	4.865
2004 Klaus	413	906	0.698	0.510	5.889	4.647
2005 Klaus	470	971	0.736	0.424	6.286	4.666
2006 Klaus	398	841	0.73	0.512	5.429	4.693
2007 Klaus	406	800	0.756	0.500	5.204	4.879
2008 Klaus	429	906	0.761	0.510	5.877	4.666
2009 Klaus	418	866	0.723	0.498	5.029	4.903
2010 Klaus	418	899	0.724	0.558	4.896	4.664
2011 Klaus	438	884	0.764	0.505	5.238	4.755
2012 Klaus	462	893	0.76	0.488	5.710	4.825
2013 Klaus	486	979	0.765	0.396	6.859	4.940
2014 Zeman	396	791	0.716	0.497	5.250	4.989
2015 Zeman	381	782	0.736	0.481	5.272	5.160
2016 Zeman	450	987	0.729	0.455	5.713	4.886
2017 Zeman	618	1,512	0.717	0.457	5.202	4.898
2018 Zeman	487	1,248	0.700	0.517	4.469	4.730
2019 Zeman	665	1,625	0.734	0.500	5.473	5.014

References

Alexeev, P. M. (1998). Graphemic and syllabic length of words in text and vocabulary. *Journal of Quantitative Linguistics*, **5**(1–2): 5–12.

Altmann, G. (1978). Zur Verwendung der Quotiente in der Textanalyse. In Altmann, G. (ed), *Glottometrika 1*. Bochum: Brockmeyer, pp. 91–106.

Altmann, G. and Köhler, R. (2015). *Forms and Degrees of Repetition in Texts: Detection and Analysis*. Berlin/Munich/Boston: de Gruyter.

Berman, R. A. and Ravid, D. (2009). Becoming a literate language user: oral and written text construction across adolescence. In Olson, D. R. and Torrance, N. (eds), *The Cambridge Handbook of Literacy*. Cambridge: Cambridge University Press, pp. 92–111.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik: Sprachstatistische Untersuchungen*. Jena: Fischer.

Chen, R. and Liu, H. (2015). Ideologies of Supreme Court justices: quantitative semantic analysis of multiple opinions of "Bush v. Gore 2000". *Glottotheory*, **6**(2): 299–322.

Choi, Y. S. and Benoit, W. L. (2013). A functional analysis of the 2007 and 2012 French Presidential debates. *Journal of Intercultural Communication Research*, **42**(3), 215–27.

Čech, R. (2014). Language and ideology: quantitative thematic analysis of new year speeches given by Czechoslovak and Czech presidents (1949-2011). *Quality & Quantity*, **48**(2): 899–910.

Čech, R. (2015). Text length and the lambda frequency structure of the text. In Mikros, G. K. and Mačutek, J. (eds), *Sequences in Language and Text*. Berlin/Boston: de Gruyter, pp. 71–87.

Covington, M. A. and McFall, J. D. (2010). Cutting the Gordian Knot: the moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, **17**(2): 94–100.

David, J., Čech, R., Davidová Glogarová, J., Radková, L., and Šústková, H. (2013). *Slovo a text v historickém kontextu – perspektivy historickosémantické analýzy jazyka [Word and Text in the Historical Context – Perspectives on the Historical-Semantic Analysis of Language]*. Brno: Host.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*, **8**(1): 107–21.

Eeg-Olofsson, M. (2009). A word length regularity and its genesis. *Glottometrics*, **19**: 49–69.

Fidler, M. and Cvrček, V. (2015). A data-driven analysis of reader viewpoints: reconstructing the historical reader using keyword analysis. *Journal of Slavic Linguistics*, **23**(2): 197–239.

Garcia, A. C. (2018). Presidential campaign talk: question-answering in 'neutral informational interviews'. *Discourse & Society*, **29**(3): 256–78.

Haswell, R. H. (2000). Document improving in college writing: a longitudinal approach. *Written Communication*, **17**(3): 307–52.

Herdan, G. (1958). The relation between the dictionary distribution and the occurrence distribution of words and its importance for the study of quantitative linguistics. *Biometrika*, **45**(1–2): 222–28.

Herdan, G. (1966). *The Advanced Theory of Language as Choice and Chance*. Berlin/Heidelberg: Springer.

Jičínský, M. and Marek, J. (2017). New year's day speeches of Czech presidents: phonetic analysis and text analysis. In Saeed, K., Homenda, W., and Chaki, R. (eds), *Computer Information Systems and Industrial Management*. Cham: Springer, pp. 110–21.

Kubát, M. (2014). Moving window type-token ratio and text length. In Altmann, G., Čech, R., Mačutek, J., and Uhlířová, L. (eds), *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM-Verlag, pp. 105–13.

Kubát, M. (2016). *Kvantitativní analýza žánrů [Quantitative Analysis of Genres]*. Ostrava: Ostravská univerzita.

Kubát, M. and Čech, R. (2016). Quantitative analysis of US presidential inaugural addresses. *Glottometrics*, **34**: 14–27.

Kubát, M., Matlach, V., and Čech, R. (2014). *QUITA – Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Kubát, M. and Milička, J. (2013). Vocabulary richness in genres. *Journal of Quantitative Linguistics*, **20**(4): 339–49.

Kuznetsova, J. (2016). Modern Russian history through the new year addresses. In Kübler, S. and Dickinson, M. (eds), *Proceedings of Computer Linguistics Fest 2016*. Bloomington (IN): Indiana University, pp. 34–38. http://ceur-ws.org/Vol-1607/kuznetsova.pdf (accessed 19 July 2019).

Lim, E. T. (2004). Five trends in presidential rhetoric: an analysis of rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, **32**(2): 328–48.

Lintunen, P. and Mäkilä, M. (2014). Measuring syntactic complexity in spoken and written language: Comparing the incomparable?. *Research in Language*, **12**(4): 377–99.

Liu, F. (2012). Genre analysis of American presidential inaugural speech. *Theory and Practice in Language Studies*, **2**(11): 2407–11.

Matić, D. (2012). Ideological discourse structures in political speeches. *Komunikacija i kultura online*, **3**: 54–78.

Mitchell, D. (2015). Type-token model: a comparative study. *Journal of Quantitative Linguistics*, **22**(1): 1–25.

Neary-Sundquist, C. A. (2017). Syntactic complexity at multiple proficiency levels of LS German speech. *International Journal of Applied Linguistics*, **27**(1): 242–62.

Popel, M. and Žabokrtský, Z. (2010). TectoMT: modular NLP framework. In Loftsson H., Rögnvaldsson E., Helgadóttir S. (eds), *International Conference on Natural Language Processing*. Berlin/Heidelberg: Springer, pp. 293–304.

Popescu, I.-I., Altmann, G., Grzybek, P. et al. (2009). *Word Frequency Studies*. Berlin/New York, NY: de Gruyter.

Popescu, I.-I., Čech, R., and Altmann, G. (2013). Descriptivity in Slovak lyrics. *Glottotheory*, **4**(1): 92–104.

Popescu, I.-I., Best, K.-H., and Altmann, G. (2014a). *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Čech, R., and Altmann, G. (2014b). Descriptivity in special texts. *Glottometrics*, **29**: 70–80.

Popescu, I.-I., Lupea, M., Tatar, D., and Altmann, G. (2015). *Quantitative Analysis of Poetic Texts*. Berlin/Boston: de Gruyter.

Quam, J. and Ryshina-Pankova, M. (2016). "Let me tell you . . .": audience engagement in the campaign speeches of Trump, Clinton, and Sanders. *Russian Journal of Linguistics*, **20**(4): 140–60.

Roitman, M. (2014). Presidential candidates' ethos of credibility: the case of the presidential pronoun I in the 2012 Hollande–Sarkozy debate. *Discourse & Society*, **25**(6): 741–65.

Savoy, J. (2010). Lexical analysis of US political speeches. *Journal of Quantitative Linguistics*, **17**(2): 123–41.

Savoy, J. (2015a). Authorship attribution using political speeches. In Tuzzi, A., Benešová, M., and Mačutek, J. (eds), *Recent Contributions to Quantitative Linguistics*. Berlin/Boston: de Gruyter, pp. 153–64.

Savoy, J. (2015b). Text clustering: an application with the State of the Union addresses. *Journal of the Association for Information Science and Technology*, **66**(8): 1645–54.

Savoy, J. (2015c). Vocabulary growth study: an example with the state of the union addresses. *Journal of Quantitative Linguistics*, **22**(4): 289–310.

Savoy, J. (2016). Text representation strategies: an example with the State of the Union addresses. *Journal of the Association for Information Science and Technology*, **67**(8): 1858–70.

Savoy, J. (2017). Analysis of the style and the rhetoric of the American presidents over two centuries. *Glottometrics*, **38**: 55–76.

Savoy, J. (2018a). Analysis of the style and rhetoric of the 2016 US presidential primaries. *Digital Scholarship in the Humanities*, **33**(1): 143–59.

Savoy, J. (2018b). Trump's and Clinton's style and rhetoric during the 2016 presidential election. *Journal of Quantitative Linguistics*, **25**(2): 168–89.

Short, D. (1993). Czech. In Comrie, B. and Corbett, G. G. (eds), *The Slavonic Languages*. London/New York, NY: Routledge, pp. 455–532.

Těšitelová, M. (1992). *Quantitative Linguistics*. Amsterdam/Philadelphia, PA: Benjamins.

Tuzzi, A., Popescu, I.-I., and Altmann, G. (2010). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM-Verlag.

Van Noije, L. and Hijmans, E. (2005). National identity and nationalism in new year's speeches of French presidents. *Communications*, **30**(1): 23–54.

Wang, Y. and Liu, H. (2018). Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 election. *Discourse & Society*, **29**(3): 299–323.

Wimmer, G. (2005). The type-token relation. In Köhler, R., Altmann, G., and Piotrowski, R. G. (eds), *Quantitative Linguistics. An International Handbook*. Berlin/New York, NY: de Gruyter, pp. 361–68.

Zörnig, P., Stachowski, K., Popescu, I.-I. et al. (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences*. Lüdenscheid: RAM-Verlag.