Does an author leave a syntactic footprint?

Radek Čech[1], Miroslav Kubát[1], Ján Mačutek[2,3], Michaela Koščová[2]

[1]University of Ostrava – cechradek@gmail.com, miroslav.kubat@gmail.com

[2]Mathematical Institute, Slovak Academy of Sciences – jmacutek@yahoo.com, kosc.michaela@gmail.com

[3]Constantine the Philosopher University in Nitra

Abstract

Results of an authorship analysis of Czech texts based on syntactic relations are presented. For the analysis, 82 annual speeches of Czech and Czechoslovak presidents delivered on the occasion of the end of the year were used. The texts were syntactically annotated according to the Universal Dependencies framework. Relative frequencies of particular syntactic relations were used to determine distances among the texts. The distances were calculated by the Cosine Delta method. Then, a hierarchical cluster analysis was performed. The results show that the method can contribute to the identification of particular presidents.

Keywords: authorship, syntax, president's speech

1. Introduction

Writing a text (or producing a speech) involves, among others, selecting the appropriate linguistic means that are used to express the intended content. Obviously, many different factors influence the selection, such as e.g. age, education, gender, the goal of writing/speaking and writer's/speaker's individuality. All of these factors leave a "footprint" in the text or speech. The main goal of stylometry to identify these "footprints". In stylometry (and in authorship analysis, its biggest domain), many approaches focusing on different language properties have been developed to date (for an overview, see Neal et al., 2018). However, only a relatively small number of them concentrate on syntax (e.g.; Baayen et al., 1996; Sidorov et al., 2014; Soler-Company and Wanner, 2016). It is not surprising especially if one realizes that until relatively recently it was not easy to obtain a reliable automatic text annotation at the syntactic level. Moreover, it is an open question to which degree syntax can mirror the authorship.

In this study, we examine whether relative frequencies of syntactic relations, as they are defined within the Universal Dependency approach (De Marneffe et al., 2021), reflect the authorship and whether they can be used for authorship analysis. We applied the method to language material consisting of a sample of Czech and Czechoslovak presidents delivered on the occasion of the end of the year (see Section 2). We chose this material because the previous analysis based on the lexical characteristics of the text (cf. Kubát et al., 2021) brought very good results with respect to the authorship identification in this language sample. The main goal of our study is to find out whether particular authors leave beside the lexical footprint also a syntactic one.

2. Language material and methodology

For the analysis, we used a corpus consisting of 82 annual speeches of chosen Czechoslovak and Czech presidents.[1] The speeches were delivered at the end or beginning of the year. They usually summarize important events from the past year and gave an outlook for the upcoming year. The authorship analysis of presidential speeches faces several problems, which should be mentioned here. First, it is known that some presidents are real authors of their speeches (e.g. Václav Havel and Václav Klaus), while speeches of some others were written in cooperation with their collaborators. From this point of view, the authorship analysis of these texts may be seen as questionable. However, each president is responsible for his speech and he hardly be expected to make a speech that he has not read beforehand and with which he disagrees (including formal characteristics of the text). Thus, in our approach, the notion of "authorship" expresses taking the political responsibility for the text.[2]

Because it was shown that there are considerable differences between the style of speeches from the communist era, on the one hand, and the democratic era, on the other, (cf. Kubát et al., 2021, where also the historical context is presented) we decided to split the original sample into two parts and analyze them separately.[3] The first sample consists of speeches given by democratic presidents (it contains 64,530 tokens), while the second sample consists of speeches given by communist presidents (it contains 70,787 tokens). Basic characteristics of the corpus are provided in Kubát et al. (2021).[4]

The texts were syntactically annotated according to the Universal Dependencies framework (https://universaldependencies.org). Based on this annotation, each text was transformed into the sequence of syntactic relations, such as e.g. nominal subject, object, adverbial modifier (see https://universaldependencies.org/u/dep/index.html for the complete list). For illustration, we present an example of the transformation of the first sentence taken from Havel's speech given in 1995:

"Milí spoluobčané, když jsem vás před pěti lety v tento den poprvé oslovil, řekl jsem, že do našich srdcí se opět vrátila naděje."[5]

"amod nsubj mark aux obj case nummod obl case det obl obl advcl root aux mark case det obl expl:pv advmod ccomp obj"

Then, relative frequencies of syntactic relations were used to evaluate distances among particular texts. The distances were calculated by the Cosine Delta method (Smith & Aldridge 2011). The choice of the method was motivated pragmatically, it gives better results than other methods used. It is also ranked among the most popular ones (probably because it brings the best results in analyses of this kind). Finally, a hierarchical cluster analysis was performed. The analysis was executed by R package Stylo (Eder et al., 2016).

[1] We omitted Masaryk's speech because he delivered the only one. Hácha's speeches were not included in this research because he cannot be assigned to either of the groups consisting of communist or democratic presidents.

[2] A similar approach was used by Tuzzi et al. (2010)

[3] We also note that results are much worse (no clear pattern for speeches by particular presidents is visible) if the corpus is not split into these two parts.

[4] For details, see https://cechradek.cz/data/cech_et_al_syntactic_footprint_list_of_texts.pdf

[5] "Dear fellow citizens, when I first addressed you on this day five years ago, I said that hope had returned to our hearts." (translated by the authors of this paper).

For calculation of distances, we use 22 most frequent syntactic relations in the case of democratic addresses and 18 most frequent syntactic relations in the case of communist addresses. The number of chosen syntactic relations was selected as follows. We checked results based on all possible configurations (number of the most frequent units, increment, culling, n-gram size) and chose the one that gave the best results for each sample separately (except for the number of the most frequent syntactic relations, we used the same settings for both samples: increment = 0, culling = 0, n-gram size = 1). We are aware that the approach is not general, but heuristic only. However, given that the primary goal of the study is to find out if the syntactic relations can be used for the authorship analysis, this approach is justifiable, in our opinion.

3. Results

Results, presented in Figures 1 and 2, show that it is possible to identify some tendencies in clustering which correspond to the authorship. One can determine four clusters among the democratic presidents (Figure 1). The most homogeneous cluster (concerning the authorship) comprises speeches given by Havel (at the bottom of the tree). However, it does not cover all of Havel's speeches (only 8 out of 14). A compelling result represents the cluster containing Beneš's speeches where we can find all his addresses, plus one Klaus gave. Similarly, we can evaluate the cluster comprising Klaus' speeches where are 9 out of 10 his total, plus one Havel's speech. The last cluster contains all of Zeman's speeches, plus four given by Havel. To sum up, in the case of both Beneš and Zeman, all their speeches fall to particular clusters. It is also similar in the case of Klaus (with one exception). Although Havel appeared in three different clusters, we can trace a clear tendency to form a common cluster in at least some of his speeches.

The analysis of communist presidents does not bring such directly interpretable results. Nevertheless, some clear tendencies are visible. First, there are 10 out of 11 speeches in the Novotný's cluster, including one of Gottwald's addresses. Next, all of Zápotocký's speeches fall to the single cluster, but its homogeneity is disrupted by three of Husák's addresses. All other Husák's speeches (11 out of 14) create a separate cluster, filled by three of Svoboda's speeches and one of Zápotocký's. We can also find Gottwald's cluster, where are 3 out of 5 speeches, and Svoboda's one, where are 3 out of 6 speeches.

4. Discussion and conclusion

Several issues should be discussed before we present the conclusions. First, this analysis's inventory of syntactic relations is relatively small (it consists of 38 relations). Using all these relations (their relative frequencies) did not bring satisfactory results. Therefore, we determined the number of syntactic relations empirically for each sample (cf. Section 2). We found out that the results are very sensitive to the number of used relations. Even minor changes in this parameter led to different clustering, although some major tendencies remained evident for all settings. It is why we used different numbers of relations for the two samples.

Second, proportions of syntactic relations reflect only a part of the syntax used by the author. Even if the same (or very similar) proportions appear, other syntactic characteristics can be quite different (e.g., the complexity of the syntactic tree or sentence length). If the aim is to analyze the "syntactic footprint" of the author from a more general point of view, a combination of several syntactic characteristics is necessary. On the other hand, the results presented in this study show clear tendencies in author clustering, even though they are based only on proportions of syntactic relations. Thus, it seems reasonable to assume the existence of a specific author's "syntactic style".

Third, in this study we introduced a specific approach to the authorship ("authorship" as the political responsibility for the text) caused by the unique nature of political speeches. In particular cases, the issue of authorship is sometimes more complicated. For instance, we can see a relatively high dispersion of Svoboda's speeches. However, if we consider the president's deteriorating health (he suffered a stroke during his presidency, was later in a coma, etc.), the dispersion is not surprising. Among the democratic presidents, Havel's speeches are the most dispersed. One of possible explanations is the fact that he was a dramatist, and as such he can be supposed to display a higher variability in style.

To sum up, the results show that the method can contribute to the identification of particular presidents. Although the text clustering using the most frequent lemmas (Kubát et al., 2021) is more accurate, syntactic relations are nevertheless surprisingly efficient, especially if one considers the limited number of syntactic functions. It is also possible to enhance authorship attribution accuracy by combining these two approaches.

Acknowledgement

Miroslav Kubát and Radek Čech were supported by the Czech Science Foundation (GAČR), project GA22-20632S, Quantitative Syntactic Stylistics of Contemporary Written Czech. Ján Mačutek and Michaela Koščová were supported by the project VEGA 2/0096/21.

References

Baayen, H., Van Halteren, H. and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3): 121-132.

De Marneffe, M. C., Manning, C. D., Nivre, J. and Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, 47(2): 255-308.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*, 8(1): 107-121.

Kubát M., Mačutek J. and Čech R. (2021). Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*, 36(1): 138-152.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y. and Woodard, D. (2018). Surveying Stylometry Techniques and Applications. *ACM Computing Surveys*, 50(6): 1–36.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A. and Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3): 853-860.

Smith P., Aldridge W. (2011). Improving authorship attribution: Optimizing Burrows' delta method. *Journal of Quantitative Linguistics*, 18(1), pp. 63-88.

Soler-Company, J. and Wanner, L. (2016). Authorship attribution using syntactic dependencies. In Angulo, C. and Godo, L., editors, *Artificial Intelligence Research and Development*, pp. 303-308. IOS Press.

Tuzzi, A., Popescu, I. I. and Altmann, G. (2010). *Quantitative analysis of Italian texts*. RAM-Verlag.

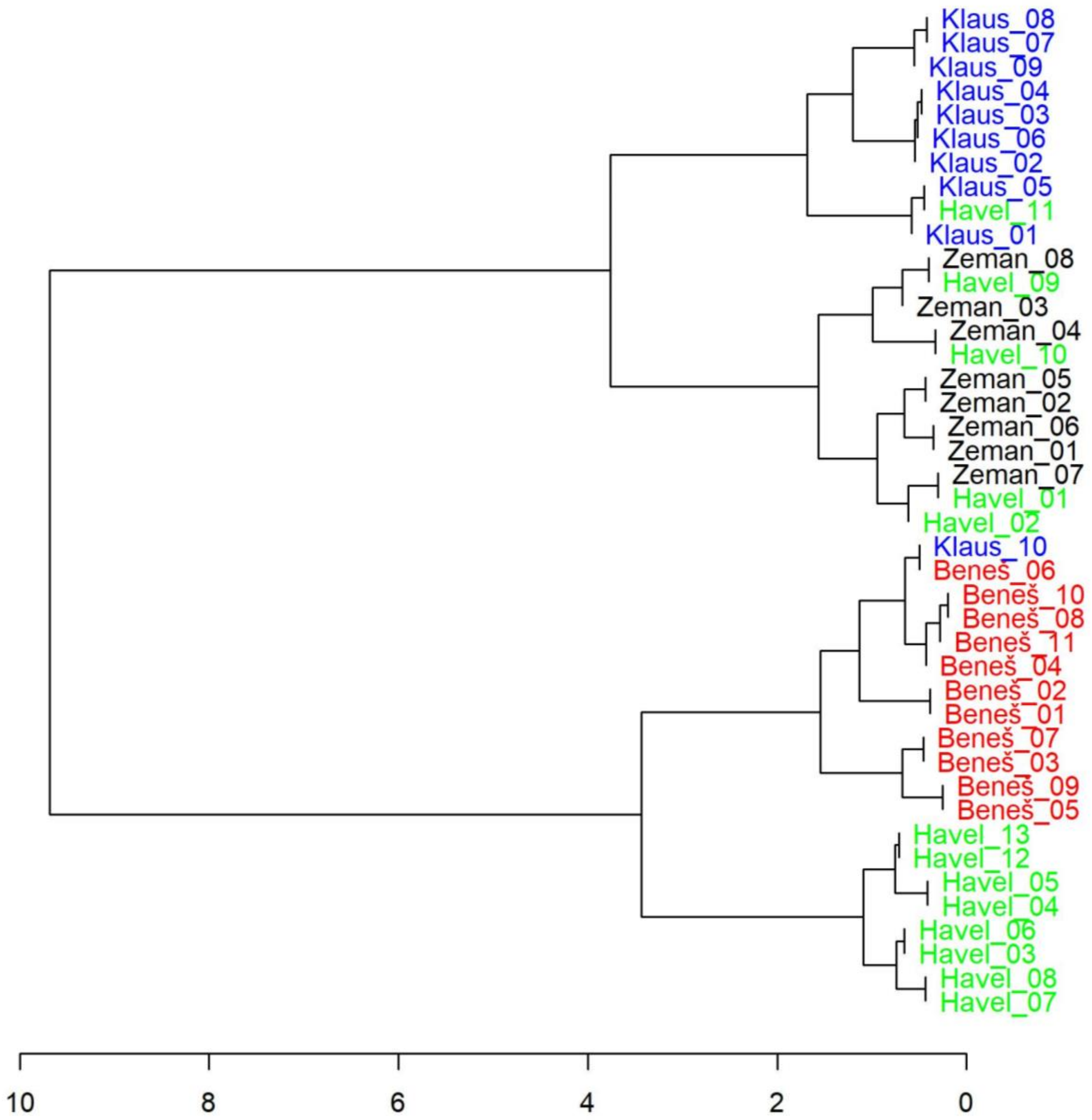

Figure 1. Cluster analysis of annual addresses given by democratic presidents.

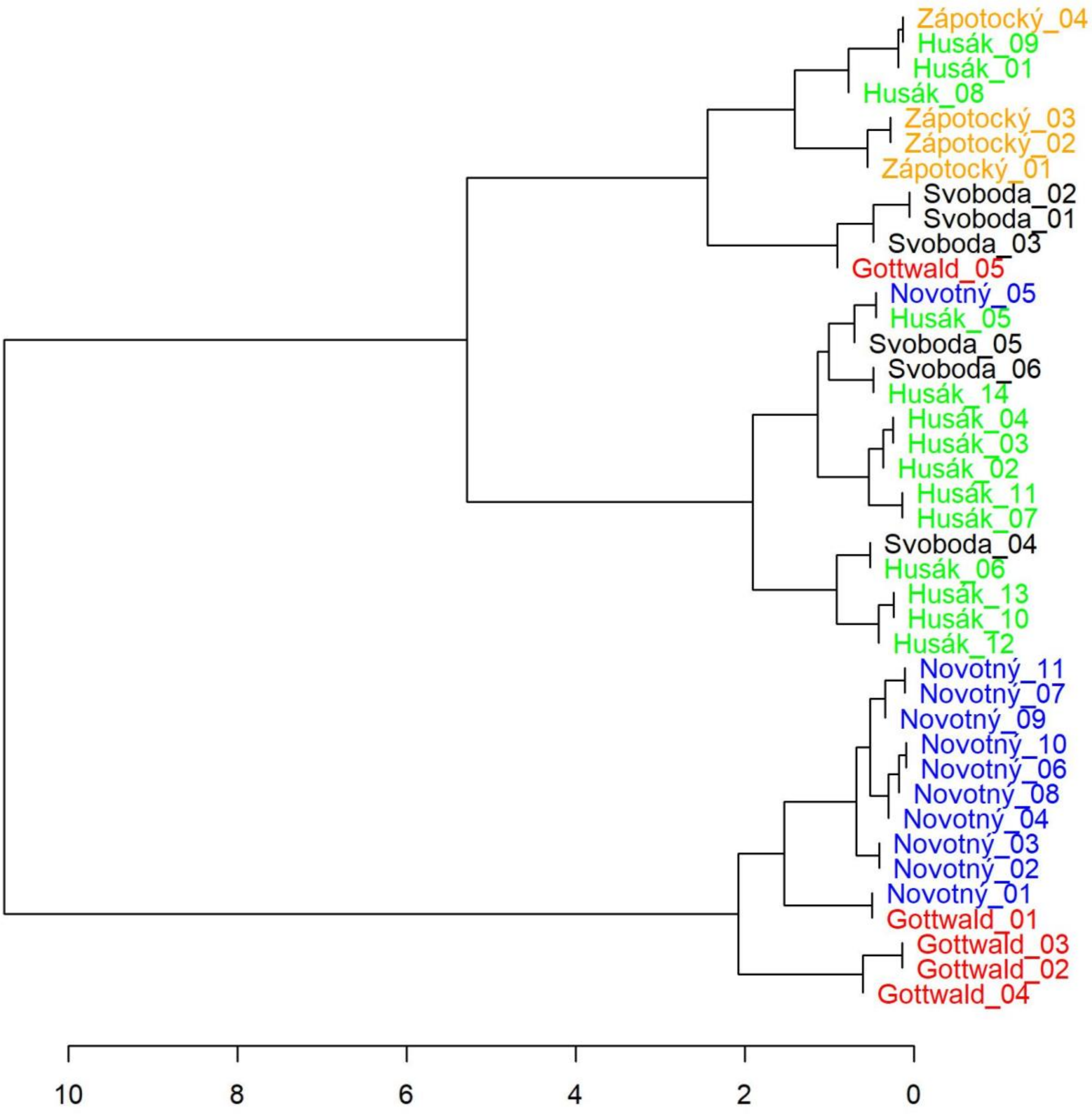

Figure 2. Cluster analysis of annual addresses given by communist presidents.