Syntactic Complexity Across Genres in Karel Čapek's Writing

Michaela Nogolová[1*], Xinying Chen[1], Miroslav Kubát[1], Žaneta Stiborská[1]

[1] University of Ostrava
[*] Corresponding author's email: nogolovam@gmail.com

DOI: https://doi.org/10.53482/2026_60_433

ABSTRACT

This paper investigates syntactic complexity across eight genres in the works of Czech writer Karel Čapek using a stylometric approach. The aim of the study is to determine whether syntactic complexity differs among the following genres: novels, short stories, travelogues, poems, newspaper columns, academic essays, children's literature, and personal correspondence. For the analysis, we compute five core complexity metrics: Average Sentence Length (in words and in clauses), Average Clause Length, Mean Dependency Distance (MDD), and Mean Hierarchical Distance (MHD). The results indicate systematic genre-specific differences in syntactic patterning within Čapek's literary work. Academic literature and travelogues consistently occupy the upper end of the complexity scale, whereas novels and short stories show comparatively low complexity, in line with their narrative and accessibility-oriented functions. Poetry is characterised by relatively low average complexity but high variability, and children's literature emerges as unexpectedly complex across several metrics. Personal correspondence and newspaper columns take up intermediate positions. Overall, the findings demonstrate that syntactic complexity offers a useful structural signal for distinguishing genres within a single author and underscore the value of syntactic features in stylometric and genre-oriented research.

Keywords: syntactic complexity, dependency syntax, genre, Karel Čapek

1 Introduction

Quantitative analysis of literary style, or stylometry, has repeatedly shown that linguistic features can reliably distinguish authors, periods and genres (e.g. Eisen et al., 2016; Grieve, 2023). Early work in computational authorship attribution and style analysis predominantly focused on lexical indicators such as word frequencies and function-word profiles (cf. Kestemont, 2014). However, over time stylometric research has increasingly moved beyond purely lexical indicators toward richer feature spaces that also incorporate syntactic, morphosyntactic, and discourse-level variables. (see Herrmann et al., 2021 for an overview). This development reflects the view that syntactic patterns, by capturing aspects of style that are less sensitive to topical variation than vocabulary alone, can provide a robust and theoretically motivated complement to lexical analysis.

The present study focuses on the works of Karel Čapek, one of the most prominent Czech authors of the twentieth century. Čapek's writing is unusually diverse in terms of genre and communicative setting: it includes novels, short stories, travelogues, poems, newspaper articles, academic literature, children's literature and personal correspondence. This makes him an ideal test case for examining the extent to which genre predicts syntactic complexity within a single author's production. Rather than asking whether syntactic complexity can distinguish between different authors, we ask how far syntactic complexity varies within an author as he moves across genres. This intra-author perspective complements more common cross-author stylometric designs and allows us to control for many confounding factors related to individual style. At the same time, we are aware that our empirical findings are restricted to Čapek's texts and do not directly extend to analysed genres in general. The study also builds on previous quantitative work on Čapek's writing, which has so far concentrated mainly on lexical and morphological characteristics (e.g., Kubát, 2016; Čech and Kubát, 2018)

Our analysis focuses on syntactic features. Using automatically parsed dependency trees, we compute five complexity metrics for each text: Average Sentence Length (in words), Average Sentence Length (in clauses), Average Clause Length (in words), Mean Dependency Distance (MDD) and Mean Hierarchical Distance (MHD). Together, these metrics capture both traditional length-based properties and dependency-based structural characteristics.

By integrating detailed syntactic analysis with a genre-focused perspective, the study aims to refine our understanding of Čapek's stylistic versatility and to contribute more generally to research on the relationship between syntax and genre. The findings not only reveal genre-specific patterns in Čapek's writing but also offer a complementary perspective to genre studies based on large, multi-author corpora (cf. Wang and Liu, 2017; Chen and Kubát, 2024).

2 Language Material and Methodology

The corpus used in this study consists of over 700 texts written by Karel Čapek, spanning eight distinct genres: novels, short stories, travelogues, newspaper columns, academic literature, poetry, children's literature, and correspondence. Due to substantial differences in text lengths across genres, the study adopts a standardized approach to define a "text" as a unit. A text is defined as one poem, one short story from a collection, one chapter from a novel, one letter of correspondence, one chapter from a scientific book, one chapter from a travelogue, or one story from a children's book. This approach ensures comparability across genres, provides a more robust dataset for analysing intra-genre variations and enabling statistical evaluation. An overview of the corpus is presented in Table 1.

Table 1: Overview of the language material.

Genres	Number of texts	Total size in tokens	Average size in tokens
novel	252	219078	869.357
short story	71	96318	1356.592
travelogue	132	49508	377.924
newspaper column	92	39042	424.370
academic literature	70	44518	635.971
poem	24	1984	82.667
children literature	26	25944	997.846
correspondence	93	26275	282.527
total	**760**	**502667**	

The syntactic analysis in this study was conducted in following steps. First, all texts were parsed using UDPipe 2.0 (Straka, 2018), employing pre-trained models from the Universal Dependencies (UD) version 2.15 dataset (Zeman et al., 2024). Following the initial parsing, the output was systematically converted into the Surface Syntactic Universal Dependencies (SUD) framework (Gerdes et al., 2018), which emphasizes syntactic functions over content-focus features and is particularly well-suited for assessing structural sentence complexity.

For data consistency, only sentences that (i) have a finite verb or auxiliary as their syntactic root, and (ii) are free of tokens containing abbreviations, numerical digits, or non-standard characters, were retained. This filtering step reduces potential noise in the dependency trees and supports more reliable computation of syntactic indices.

In this analysis, we use Average Sentence Length (ASL), Average Clause Length (ACL), Mean Dependency Distance (MDD), and Mean Hierarchical Distance (MHD). Together, these indicators capture both surface-level and deep structural characteristics of syntax.

ASL was calculated using two complementary methods: one based on the word count per sentence, and another based on the number of clauses per sentence. While the first method captures overall sentence elaboration, the second reflects the density of subordinate structures within sentences.

ACL was derived by dividing the total word count by the total number of clauses, offering insight into the internal complexity of clauses themselves.

MDD, as formulated by Liu (2008), assesses the average linear distance between a dependent and its syntactic head. For each word in the text (excluding root tokens and punctuation), the dependency distance was calculated as the absolute difference between the word's index and that of its syntactic parent. The average was then computed as follows:

$$MDD = \frac{\sum_{i=1}^{n-s} |DD_i|}{n-s}$$

where n denotes the total number of words, s is the number of analysed sentences and DD represents the dependency distance of the i^{th} word. Because the root of the sentence is excluded from the calculation, it must be omitted from the total word count; hence, $n - s$.

MHD, proposed by Jing and Liu (2015), offers a hierarchical perspective by measuring the average vertical depth of dependency trees. For each token[1] in the text, the hierarchical distance (HD) is defined as the number of dependency links from the token to the root of the sentence. The average value of these distances provides a measure of how deeply nested syntactic structures are, serving as a proxy for syntactic embedding.

The differences between genres were statistically tested. For each pairwise comparison, we first assessed the normality of the distributions using the Shapiro–Wilk test (Shapiro and Wilk, 1965). If both groups met the normality assumption, we compared their means with an independent-samples t-test. If at least one group deviated from normality, we used the Mann–Whitney U test (Mann and Whitney, 1947) as a non-parametric alternative. Since multiple pairwise comparisons were carried out, we adjusted the resulting p-values for multiple testing using the Benjamini–Hochberg correction (Benjamini and Hochberg, 1995).

3 Results

This section presents the results of the syntactic complexity analysis of Karel Čapek's multi-genre corpus. Focusing first on Average Sentence Length (ASL) in words, based on the results of statistical tests (see Table 2), the genres fall into four distinguishable groups (see Figure 1). The first group, with the lowest[2] ASL, comprises poetry, novels and short stories. Poetry shows both the shortest sentences and the highest variability, which is expected given that poetic syntax must adapt to metrical and prosodic constraints, alternating between very short lines and more extended sentences. Novels and short stories also favour relatively short sentences, in line with their communicative goal of maintaining narrative flow and ensuring processing ease for a broad readership.

A second group is formed by personal correspondence and children's literature, which exhibit intermediate – and in Čapek's case noticeably higher – sentence lengths. In his letters, Čapek writes to specific, often highly educated addressees, including the first Czechoslovak president Tomáš Garrigue Masaryk and the poet, literary critic, journalist and translator S. K. Neumann. In this relatively unconstrained context, he allows himself longer, more expansive sentences, which is reflected in the higher ASL values. The relatively long sentences in children's literature are less intuitive, given the usual expectation of shorter, simpler sentences in texts for young readers. Here, the patterns suggest that Čapek does not

[1] Except for the root and punctuation which are excluded.
[2] In all figures, the boxplots are ordered by their median values, as most distributions deviate from normality and the median therefore provides a more appropriate basis for comparison than the mean.

substantially simplify his syntax for children but relies on other strategies (such as topic choice or lexical transparency) to ensure comprehensibility.

Newspaper articles constitute a separate, third group: their ASL is significantly different from all other genres, and they occupy a position just below the most complex group. This is consistent with the genre's dual function. On the one hand, newspaper columns are information-dense and argumentative, which encourages relatively long, multi-clause sentences; on the other hand, they still address a broad, non-specialist audience, which may limit syntactic complexity compared to strictly academic prose.

The fourth group, with the highest ASL values, consists of academic prose and travelogues. Both genres are typically expository and information-heavy: academic texts condense complex arguments and theoretical discussion into syntactically dense sentences, while travelogues frequently contain extended descriptive passages and background explanations.

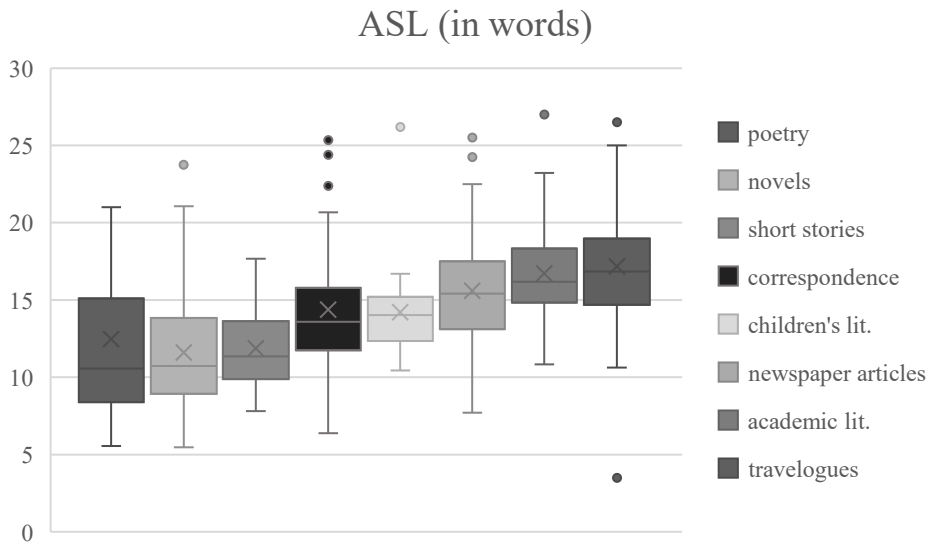

Figure 1: Results of Average Sentence Length in the number of words.

Table 2: Statistical differences (p-value) between genres based on ASL (in words).

Genres	Poetry	Novels	Short stories	Correspondence	Children's lit.	Newspaper articles	Academic lit.
novels	0.745						
short stories	0.264	0.190					
correspondence	**0.003**	**<0.001**	**<0.001**				
children's lit.	**0.007**	**<0.001**	**<0.001**	0.819			
newspaper articles	**<0.001**	**<0.001**	**<0.001**	**0.008**	**0.029**		
academic lit.	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.014**	
travelogues	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.003**	0.599

Turning to Average Sentence Length in clauses (ASL in clauses; see Figure 2 and Table 3) and Average Clause Length in words (ACL; see Figure 3 and Table 4), we find a broadly similar genre hierarchy as for ASL in words, but with some informative shifts in how complexity is realised.

Poetry has the lowest ASL in clauses and one of the lowest ACL values. It also shows considerable variability in ASL in clauses, so that it is not statistically different from most other genres (academic prose, novels, travelogues, short stories and newspapers). This pattern suggests that Čapek flexibly adjusts the number of clauses per sentence in poetry, while keeping the clauses themselves relatively short and fairly stable in length.

Academic prose has the second-lowest median of ASL in clauses (slightly above two clauses per sentence), which might initially seem at odds with its status as a complex genre. However, it exhibits the longest statistically significant ACL of all genres. Rather than piling up large numbers of clauses, Čapek tends to pack a great deal of information into individual clauses, so that much of the complexity resides within clauses rather than in long clause chains. Travelogues show a related pattern: they have somewhat more clauses per sentence than academic literature, but still relatively moderate ASL in clauses and the second-highest ACL. Again, complexity is driven by rich, information-dense clauses rather than by extreme clause stacking.

Novels and short stories have both relatively low ASL in clauses and relatively short clauses. This combination points to syntactically lighter sentences overall, consistent with genres that prioritise readability and narrative flow and avoid very complex structures, both in terms of how many clauses are combined and how long those clauses are. Personal correspondence, by contrast, features more clauses per sentence on average, while clause length remains in the mid-range. Here, Čapek appears to allow himself more expansive sentences made up of several medium-length clauses, which fits the more informal and personalised communicative setting.

Newspaper articles occupy an upper-middle position: they have the third-highest number of clauses per sentence and the third-longest clauses. They thus combine both types of complexity – more clauses and longer ones – but still remain below the most extreme values observed in the corpus. The most striking pattern is found in children's literature, which has the highest ASL in clauses and medium clause lengths. In other words, Čapek builds unusually long clause chains even when writing for children and does not fully reduce clause-based complexity.

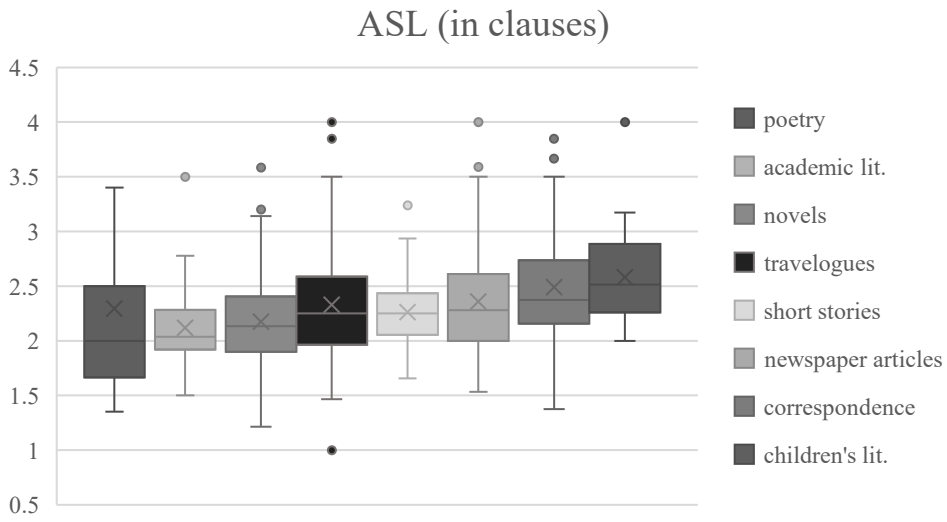

Figure 2: Results of Average Sentence Length in the number of clauses.

Table 3: Statistical differences (p-value) between genres based on ASL (in clauses).

Genres	Poetry	Academic lit.	Novels	Travelogues	Short stories	Newspaper articles	Correspondence
academic lit.	0.597						
novels	0.394	0.252					
travelogues	0.116	**<0.050**	**<0.050**				
short stories	0.130	**<0.050**	**<0.050**	0.900			
newspaper articles	0.085	**0.001**	**<0.050**	0.628	0.541		
correspondence	**<0.050**	**<0.001**	**<0.001**	**<0.050**	**<0.050**	0.054	
children's lit.	**<0.050**	**<0.001**	**<0.001**	**<0.050**	**0.001**	**<0.050**	0.230

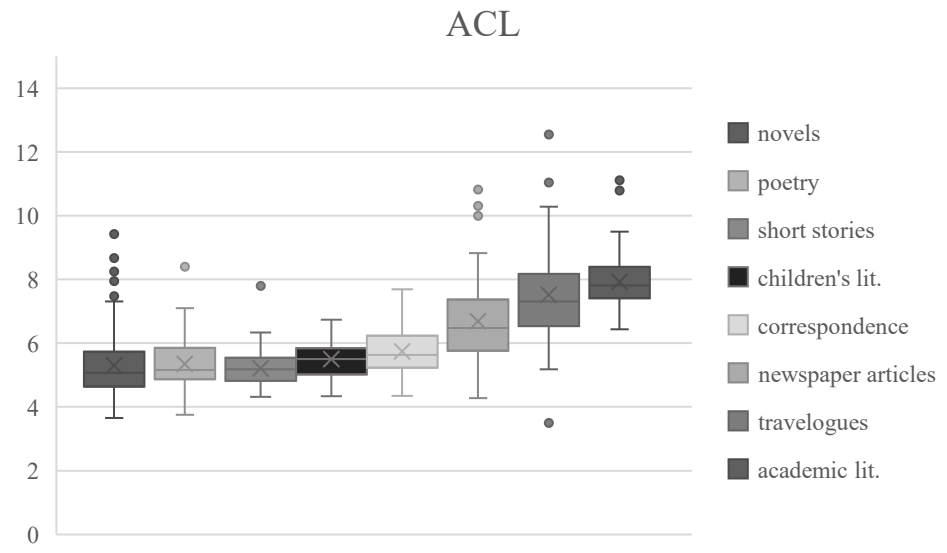

Figure 3: Results of Average Clause Length.

Table 4: Statistical differences (p-value) between genres based on ACL.

Genres	Novels	Poetry	Short stories	Children's lit.	Correspondence	Newspaper articles	Travelogues
poetry	0.584						
short stories	0.691	0.616					
children's lit.	**<0.050**	0.190	**<0.050**				
correspondence	**<0.001**	**<0.050**	**<0.001**	0.182			
newspaper articles	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**		
travelogues	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	
academic lit.	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

For Mean Dependency Distance (MDD), we again find statistically significant differences across genres (see Figure 4 and Table 5), although the overall effect is relatively modest. The clearest contrasts involve novels and short stories on the one hand and children's literature on the other. Novels and short stories have the lowest MDD values, whereas children's literature – somewhat unexpectedly – shows the highest MDD. Children's literature differs significantly from most other genres except correspondence, poetry and travelogues, while novels are significantly lower than all of them. Short stories occupy a slightly higher position than novels and do not differ significantly from poetry or academic prose. Poetry once more exhibits substantial variability but does not systematically occupy either extreme. The generally compressed range of MDD values is in line with Futrell et al. (2015), who argue that language users tend to minimise dependency distance due to working-memory constraints. In Čapek's writing, this suggests that although he clearly differentiates genres through sentence length and clause structure, genre-specific choices are constrained by a more general pressure to keep dependency distances within manageable limits.

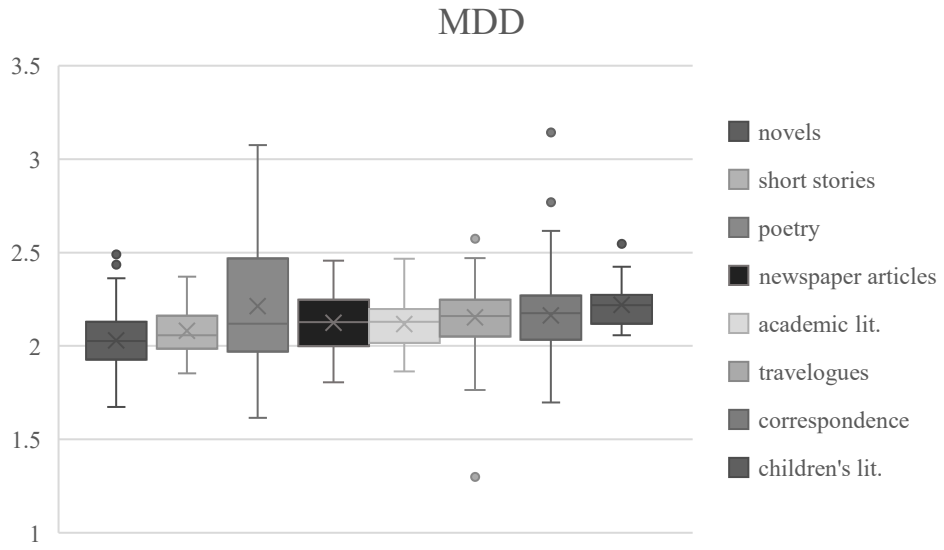

Figure 4: Results of Mean Dependency Distance.

Table 5: Statistical differences (p-value) between genres based on MDD.

Genres	novels	short stories	poetry	newspaper articles	academic lit.	travelogues	correspondence
short stories	**<0.050**						
poetry	**<0.050**	0.267					
newspaper articles	**<0.001**	**<0.050**	0.082				
academic lit.	**<0.001**	0.075	0.074	0.751			
travelogues	**<0.001**	**<0.001**	0.909	0.162	0.051		
correspondence	**<0.001**	**<0.050**	0.906	0.238	0.090	0.900	
children's lit.	**<0.001**	**<0.001**	0.937	**<0.050**	**<0.001**	0.060	0.094

Finally, Mean Hierarchical Distance (MHD) reveals a relatively clear separation between individual genres (see Figure 5). According to the statistical tests, we can again identify four distinct groups (see Table 6). The first group is formed by poetry, which has the lowest MHD. This is in line with the genre's tendency towards structurally flat, often elliptical sentences, in which many relations are left implicit rather than being expressed through deeply nested structures.

The second group comprises novels, short stories and children's literature, which show low to intermediate MHD values. Here, the syntactic structures are usually organised into moderately complex sentences: there is some hierarchical embedding, but it is kept within limits, so that the overall sentence architecture remains relatively transparent. This pattern fits genres that rely on storytelling and scene-building.

The third group, with higher MHD, consists of personal correspondence, newspaper articles and academic literature. The highest, and statistically clearly distinct MHD values are found in travelogues. This suggests that Čapek's travel writing makes particularly extensive use of embedded words in the sentence to layer descriptions, background information and commentary, creating syntactic structures that are not only long but also deeply nested in terms of their hierarchical organisation.

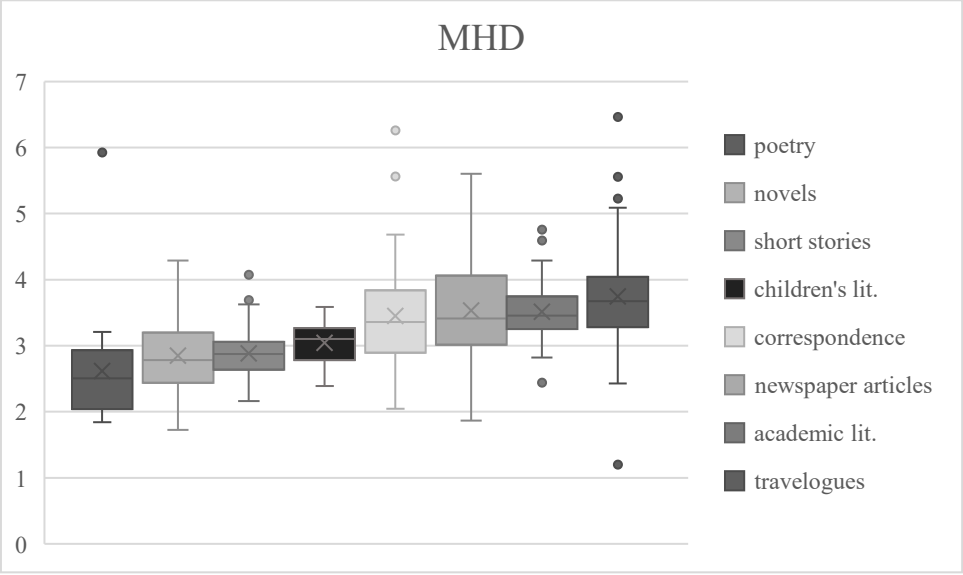

Figure 5: Results of Mean Hierarchical Distance.

Table 6: Statistical differences (p-value) between genres based on MHD.

Genres	poetry	novels	short stories	children's lit.	correspondence	newspaper articles	academic lit.
novels	**<0.050**						
short stories	**<0.001**	0.270					
children's lit.	**<0.001**	**<0.050**	0.060				
correspondence	**<0.001**	**<0.001**	**<0.001**	**<0.050**			
newspaper articles	**<0.001**	**<0.001**	**<0.001**	**0.001**	0.376		
academic lit.	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.180	0.447	
travelogues	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.001**	**<0.050**	**<0.050**

4 Conclusion

This study examined how syntactic complexity varies across eight genres in the work of a single author, Karel Čapek, using five syntactic metrics. Taken together, the results show that genre is a strong predictor of syntactic profile even when authorial style is held constant. The patterns are far from random: each genre occupies a relatively stable position across the different indices, and these positions align in meaningful ways with the communicative function and audience design of the genres.

Across the metrics, academic prose and travelogues consistently cluster at the upper end of the complexity scale. They feature relatively long clauses and, in the case of travelogues in particular, deeply nested hierarchical structures. At the other end of the spectrum, novels and short stories are systematically less complex: they have shorter sentences, fewer and shorter clauses, and lower dependency and hierarchical distances. This profile fits their role as narrative genres that prioritise readability and narrative flow. Poetry stands somewhat apart: it is typically less complex in terms of average lengths and hierarchical depth, yet it shows the highest variability in most indices. This variability reflects the flexibility of poetic form, where syntactic shaping is strongly influenced by prosodic, rhythmic and stylistic choices.

Perhaps the most striking finding concerns children's literature. Contrary to common expectations about texts for young readers, Čapek's children's books are syntactically far from simplified. They show relatively long sentences, high numbers of clauses per sentence and surprisingly high linear and hierarchical complexity. This suggests that Čapek does not primarily adapt his style for children by reducing syntactic sophistication; instead, he may rely more on other resources – such as familiar topics or transparent lexis – to support comprehension. Personal correspondence and newspaper columns occupy intermediate positions: they permit more expansive and sometimes embedded sentences than narrative prose but remain less structurally extreme than academic writing or travelogues.

A key contribution of this study lies in the demonstration of Čapek's stylistic adaptability. Rather than maintaining a uniform syntactic style across genres, he modulates his writing with a high degree of genre sensitivity. On a broader methodological level, the study illustrates the added value of syntactic analysis in stylometry. While lexical and morphological features have long dominated computational

literary studies, the findings here confirm that syntactic metrics offer unique and complementary insights. The combined use of hierarchical and linear syntactic features provides a more holistic view of stylistic variation, particularly when applied across structurally diverse genres.

In conclusion, this research not only sheds light on the genre-specific stylistic practices of Karel Čapek but also contributes to a deeper understanding of how syntax functions as a vehicle of genre differentiation. It opens up avenues for further work on cross-linguistic and multilingual corpora, comparative stylistics and the evolution of genre norms over time. Ultimately, it confirms that syntactic complexity is not merely a property of linguistic form, but a meaningful dimension of literary style, integral to the way texts communicate, persuade and resonate with their readers.

Acknowledgements

This research is supported by Grant SGS04/FF/2025, University of Ostrava.

References

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), pp. 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Čech, R., Kubát, M. (2018) Morphological Richness of Text. In Fidler, M., Cvrček, V. (eds.). Taming the Corpus. *From Inflection and Lexis to Interpretation*. Springer, pp. 63–77.

Chen, X., Kubát, M. (2024). Quantifying Syntactic Complexity in Czech Texts: An Analysis of Mean Dependency Distance and Average Sentence Length Across Genres. *Journal of Quantitative Linguistics*, 31(3), pp. 260–273, https://doi.org/10.1080/09296174.2024.2370459

Eisen, M., Ribeiro, A., Segarra, S., Egan, G. (2018). Stylometric analysis of Early Modern period English plays, *Digital Scholarship in the Humanities*, 33(3), pp. 500–528, https://doi.org/10.1093/llc/fqx059

Futrell, R., Mahowald, K., Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages, *PNAS.* 112(33), pp. 10336–10341, https://doi.org/10.1073/pnas.1502134112

Gerdes, K., Guillaume, B., Kahane, S., Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In de Marneffe, M.-C., Lynn, T., Schuster, S. *Proceedings of the Second Workshop on Universal Dependencies* (UDW 2018), Brussels, Belgium. Association for Computational Linguistics, pp. 66–74.

Grieve, J. (2023). Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1), pp 47–77. https://doi.org/10.1515/cllt-2022-0040

Herrmann, J., B., Jacobs, A., M., Piper, A. (2021). Computational Stylistics. In Kuiken, D., Jacobs, A., M. (Eds.) *Handbook of Empirical Literary Studies*. De Gruyter, pp. 451–486. https://doi.org/10.1515/9783110645958-018

Jing, Y., Liu, H. (2015). Mean Hierarchical Distance Augmenting Mean Dependency Distance. In Nivre, J., Hajičová, E. (Eds.) *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Sweden, pp. 161–170.

Kestemont, M. (2014). Function Words in Authorship Attribution. From Black Magic to Theory?. In Feldman, A., Kazantseva, A., Szpakowicz, S. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature* (CLFL), 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Kubát, M. (2016). Kvantitativní analýza žánrů. University of Ostrava.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), pp. 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4), pp. 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college‐level ESL writers' language development. *TESOL quarterly*, 45(1), pp. 36–62. https://doi.org/10.5054/tq.2011.240859

Mann, H. B., Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18, pp. 50–60.

Shapiro, S. S., Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52, pp. 591–611.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 197–207.

Wang, Y., Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59, pp. 135–147. https://doi.org/10.1016/j.langsci.2016.09.006

Zeman, D. et al. (2024). Universal Dependencies 2.15, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Prague: Charles University, Accessible at http://hdl.handle.net/11234/1-5787.