sciendo
DOI 10.2478/jazcas-2025-0033

QUANTITATIVE CORPUS ANALYSIS OF VLADIMIR PUTIN'S SPEECHES

PETR POŘÍZKA[1] – POLINA IVANKOVA[2]
[1]Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic (ORCID: 0000-0001-6980-9148)
[2]Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic

POŘÍZKA, Petr – IVANKOVA, Polina: Quantitative Corpus Analysis of Vladimir Putin's Speeches. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 366 – 377.

Abstract: This study presents a quantitative analysis of Vladimir Putin's public speeches from 1999 to 2024, utilizing a corpus of approximately 1.75 million words sourced from the official Kremlin website. Using computational linguistic techniques, including hierarchical cluster analysis, TF*IDF keyword extraction and temporal trend analysis, the research systematically examines the evolution of Putin's rhetorical style and thematic content. Significant stylistic and thematic shifts are identified, particularly around 2012, coinciding with his return to the presidency and a notable increase in authoritarian governance. Prominent themes such as national identity, economic policy, security, and international conflicts vary significantly in prominence and shift in emphasis in response to key historical and political milestones. The findings reveal clear correlations between Putin's language patterns and major political or critical events, including the Chechen conflict, Putin's political rise and return to the presidency, the annexation of Crimea, and the recent military interventions in Ukraine. These findings demonstrate Putin's strategic rhetorical adaptability within changing geopolitical contexts.

Keywords: corpus, hierarchical clustering, keyword analysis, quantitative analysis, political discourse, Vladimir Putin, TF*IDF

1 INTRODUCTION AND PREVIOUS RESEARCH

Quantitative linguistics is concerned with measurable aspects of language, and its application to political discourse offers a systematic and potentially more objective way of analysing large amounts of textual data than standard methods based on introspection (without the use of computer-based analysis). Using quantitative analysis, it is possible to find patterns and trends that might remain hidden in a purely qualitative approach. In this paper we try to find quantitatively important patterns, tendencies or differences in Putin's language. It is also the first analysis of its kind on such a large dataset, processing Putin's speeches over the last 25 years.

Previous studies of Putin's speeches have predominantly used qualitative discursive methods. These include comparative discourse analyses of speeches by Putin and Biden (Kopik 2023), rhetorical analyses using Fairclough's framework (Shahbaz and Nawab 2024), critical discourse analyses comparing statements by Putin and Zelensky during the invasion of Ukraine (Tutar and Bağ 2023), speech act theory analyses highlighting Putin's performative language (Fafiyebi 2025), and

analyses of Putin's narrative strategies during key events such as the 'special military operation' (Kadim 2023). Other studies focus on comparisons of war rhetoric (Chiluwa and Ruzaite 2024), socio-cognitive perspectives (Al-Manaseer 2025), and linguistic arguments in support of geopolitical claims, notably in Putin's 2021 essay on Ukrainian dialects (Maxwell 2025). These analyses tend to be narrow in scope, focusing on specific speeches or short periods of time.

Quantitative studies on Putin's rhetoric remain rare and tend to focus on narrowly defined topics. Only three academic papers have been identified: Wang and Zeng (2023) compare stylometric features and political themes in speeches by Putin, Medvedev, Trump, and Obama; Oleinik (2023) evaluates Cohen's d and Z-scores for term specificity in war-related political discourse; and Janda et al. (2022) apply keymorph analysis to examine changes in grammatical case usage (e.g. *Russia*, *Ukraine*, *NATO*) before and after the invasion of Ukraine.

2 CORPUS OF PUTIN'S SPEECHES

We used the speeches and texts available on the official website of the President of the Russian Federation (http://kremlin.ru/) as our data source. The total number of hours of Putin's speeches available on this website exceeds 7,500 hours; we processed and analysed approximately 4,000 hours of material. The corpus includes speeches to citizens, journalists, ambassadors, politicians, media interviews, press conference recordings, meetings with government and military officials, urgent security meetings, speeches on the occasion of holidays (Victory Day, Family Day, New Year, etc.).

The corpus covers the entire period of Putin's top political career from 1999 to 2024, the texts are in TXT format (no further annotations yet) and are bilingual: in the original Russian and in Czech translation. Because of the size of the data, the Czech translation was created automatically using the AI tool *DeepL* (https://www.deepl.com/). Given the genetic and typological similarities (Slavic languages of the inflectional type), we can assume that the translation is largely accurate. The data are segmented by years or multi-year periods (e.g. 1999–2000, 2001, 2002, ... etc. up to the 2020–2022, 2023–2024 collections). The year 2012 is divided into two sub-collections, the first up to May 2012, when Putin was still Prime Minister, and the texts after this period, which fall within his second presidential term (see below). The total size of the corpus we used for the analysis is about 1.75 million words (1,751,134 tokens).

At the moment it is a simple set of texts, but our plan is to create a standard trilingual corpus (original Russian + Czech and English translation) with linguistic annotations, which will allow a deeper and more detailed study of the linguistic aspects of Putin's speeches (including grammar).

2.1 Historical and political context

The corpus covers different political phases in Vladimir Putin's career. He was Prime Minister from 1999 to 2000 and President during his first term from 2000 to

2008. From 2008 to 2012, he was prime minister again under President Dmitry Medvedev. Since 2012, Putin has been president during a period characterised by increasing authoritarianism, centralisation of power, assertive nationalism, growing international isolation, intensified and militarised foreign policy aggression, culminating in the annexation of Crimea in 2014 and the invasion of Ukraine, key events that have set the tone and focus of his political rhetoric in recent years.

3 CLUSTER ANALYSIS OF TEXT CORPUS

Hierarchical cluster analysis is a powerful unsupervised learning method used in stylometry to uncover latent structures within high-dimensional textual data. This approach is particularly suited to diachronic investigations, allowing researchers to trace stylistic shifts over time without making prior assumptions about text categorisation (Burrows 2002; Hoover 2003; Eder et al. 2016).

The cluster analysis (run in R and the *stylo* package with default settings) used the 100 most frequent words (MFW) in each text as input variables. The data were clustered using three different distance metrics: *cosine*, *Euclidean*, and *min-max distance*.[1] These metrics were chosen to test the stability of the clustering outcomes across different models of similarity. The results are shown in Fig.1.[2]

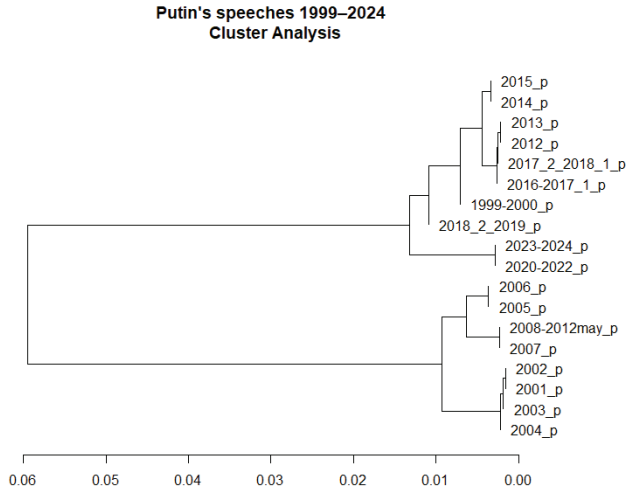

Fig. 1. Hierarchical cluster analysis of Vladimir Putin's speeches (metric: cosine; 100 most frequent words (MFW) in each text as input variables)

[1] Cosine distance measures directional similarity, highlighting relative patterns of word usage regardless of absolute frequency. Euclidean distance is sensitive to size, emphasising differences in absolute word counts. Min-max distance focuses on the range of word frequencies, making it more sensitive to outliers or rare word inflation.

[2] This is the plot with the cosine metric. All three dendrograms are available here: https://mega.nz/folder/sdtjWBbQ#vscNCLZc3kB-KT5S8jlCRw.

All three dendrograms show a coherent and converging structure, with a clear bifurcation into two primary stylistic branches, indicating a significant shift in Putin's discourse style or rhetorical strategy. Putin's political rhetoric can thus be clearly divided into pre-2012 and post-2012 stylistic periods. This chronological threshold corresponds closely with major political transitions and geopolitical events in Vladimir Putin's career. Specifically, this divide coincides with his return to the presidency in 2012, the subsequent annexation of Crimea in 2014, and the onset of increasingly centralised and confrontational governance. This shift appears to reflect a fundamental change, in line with broader political developments, including the consolidation of an increasingly autocratic regime. The high consistency of the identified clusters across different distance measures supports the reliability of the analysis.

Of interest is the position of the first speeches from 1999–2000, which appear consistently in segments of texts after 2012. In the case of the cosine and Euclidean metrics, it coincides with texts from 2012–2018, and for the min-max distance it is associated with speeches from the last years 2000–2024. Future analyses will therefore need to look more closely at possible causes. According to the dendrogram, texts from 1999–2000 can be seen as transitional or outlying observations. Such placement may reflect initial stylistic or rhetorical experimentation or variability before a consistent style is established.

4 QUANTITATIVE ANALYSIS OF KEYWORDS (TF*IDF METHOD)

The TF*IDF (Term Frequency-Inverse Document Frequency) method is a statistical measure widely used in text analysis to identify and rank significant keywords within textual data (Salton and Buckley 1988; Ramos 2003). The approach combines two aspects: term frequency (TF), which measures how often a keyword occurs in a given document or text segment, and inverse document frequency (IDF), which reflects how rarely a keyword occurs in all analysed documents or segments (Rajaraman and Ullman 2011). The resulting TF*IDF score thus assigns higher values to keywords that are characteristic of specific text periods or documents, making it particularly effective for longitudinal analyses of political discourse, such as the speeches of Vladimir Putin analysed in this study, because it reveals the main or most important expressions, motifs or themes.

In our application, keywords (hereafter KWs) were extracted and hierarchically ranked based on their TF*IDF scores across defined time slots (1999–2024), allowing for a nuanced identification of thematic shifts and continuities in political rhetoric over time.

We extracted the first 100 KW from each file using the KER tool (Libovický 2016). We then removed items that could bias the analysis, in particular proper names of persons that appeared in the interview transcript as labels preceding the dialogue line (i.e., primarily *Vladimir*, *Vladimirovich*, *Putin*, and the names of the

journalists who conducted the interview). Other names of people who were part of the text (e.g. *Yanukovych*, *Bush*, *Obama*), as well as names of institutions, places, etc., remained in the lists. We also removed the repetitive formal words *dear*, *sir*, *question*, which were also related to the format of the interview (addressing the guest). Due to the large amount of data, all extracted KWs from all files are shared in the cloud[3] and we focus mainly on their interpretation.

4.1 Thematic clusters of Putin's keywords

Total KWs: 1,612

Unique KWs: 330

The average (mean) TF*IDF score across all keywords is about 0.01066, with values ranging from about 0.00586 to 0.06022.

Based on these 330 unique keywords, we have identified 8 main thematic clusters of Putin's rhetoric:

1. National identity and geopolitics
Sample KWs: *Russia, Russian, Chechnya, federation, state, citizen, national, sovereignty, territory, international, partner, relationship, region, Donbas, Crimea, Sevastopol, Dagestan, foreign*

2. Economic and financial matters
Sample KWs: *Economy, economic, financial, market, ruble, budget, growth, investment, bank, money, enterprise, business, pension, rate*

3. Security, military, and defense
Sample KWs: *Security, military, defense, weapon, terrorism, terrorist, counter-terrorism, nuclear, war, armed, threat*

4. Political institutions and governance
Sample KWs: *Government, president, Duma, law, constitution, legal, authority, organization, council, reform, process*

5. Social policies and domestic welfare
Sample KWs: *Society, social, education, housing, healthcare, population, community, future, reform* (in a social context)*, citizen* (also appears in this group)

[3] All the extracted keywords with TF*IDF values are available here: https://mega.nz/folder/sdtjWBbQ#vscNCLZc3kB-KT5S8jlCRw. For the purposes of this study, we also translated all KWs into English using the Gemini AI tool (https://gemini.google.com/). All lists are therefore bilingual.

6. International relations and conflict
Sample KWs: *NATO, sanctions, cooperation, partnership, dialogue, threat, crisis, conflict, Bush, Obama, Yanukovych, alliance, war, global*

7. Science, technology, and innovation
Sample KWs: *Technology, technological, AI, innovation, project, modern, internet*

8. Miscellaneous and context-specific terms
Sample KWs: names (e.g. *Bush, Obama, Yanukovych*), specific events or adjectives (e.g. *clear, certain, big, real*), technical descriptors and less common or specific terms such as *medium*.

For the purposes of our analyses, we have simplified and reduced these categories to the following political themes, which indicate the main trends and changes in Putin's rhetoric. The results are summarised in the following graph (Fig. 2), which shows the thematic focus over time:

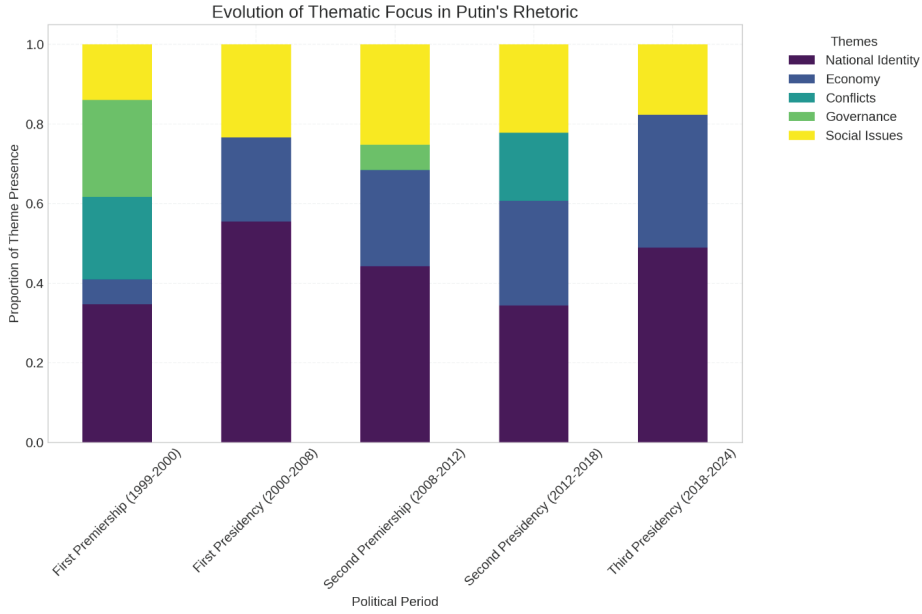

Fig. 2. Development of the thematic focus of Putin's rhetoric

National identity remains a consistently important topic (especially during the first presidential term), but its context changes – from internal identity building to Russia's international position. Conflict themes change according to the geopolitical situation: the Chechen conflict dominates the first term, while Ukraine and Crimea

appear in the second term (2012–2018). Economic themes gradually become more important, especially in the later periods. Social issues (*human*, *problem*, *work*) are consistently present, but their importance increases especially in the second premiership. Governance: domestic political topics (*Duma*, *laws*) are particularly prominent in the first premiership, but later recede into the background; international themes are particularly prominent in the second premiership (*Ukraine*, *Crimea*).

4.2 Visual analysis of the top 15 KWs and keyword correlations

The most consistent keywords throughout Putin's career reveal his core rhetorical focus: '*Russia*', '*Russian*', '*country, land*', '*person, human, man*', '*problem*', '*development*', '*state*', '*economy*', '*important*', '*work*'. Russia and Russian identity remain the absolute foundation of Putin's rhetoric across all periods, followed by references to the country/state and the Russian people.

The heat-map in Fig. 3 shows how the importance of keywords has changed over the course of Putin's career. A darker colour indicates a higher ranking (lower number = more important). Blank spaces don't mean that Putin didn't use it in a particular period, but that the word was not among the top 15 keywords in that period.

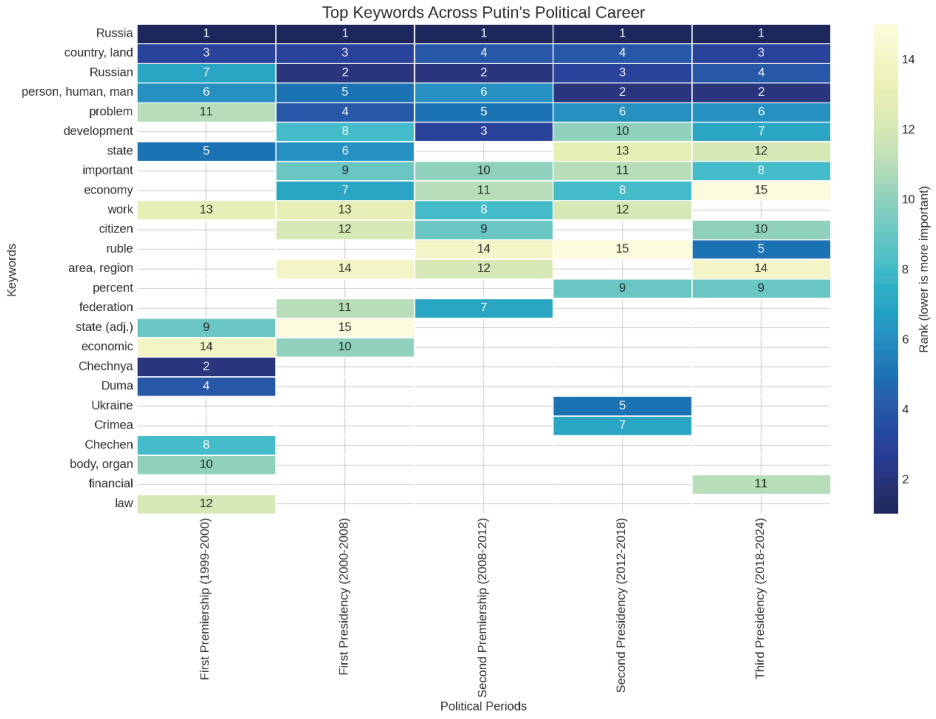

Fig. 3. Top 25 keywords from Putin's political career; blanks indicate absence from the period's top 15

The following heat-map in Fig. 4 presents a correlation map of the top 50 keywords of Vladimir Putin:

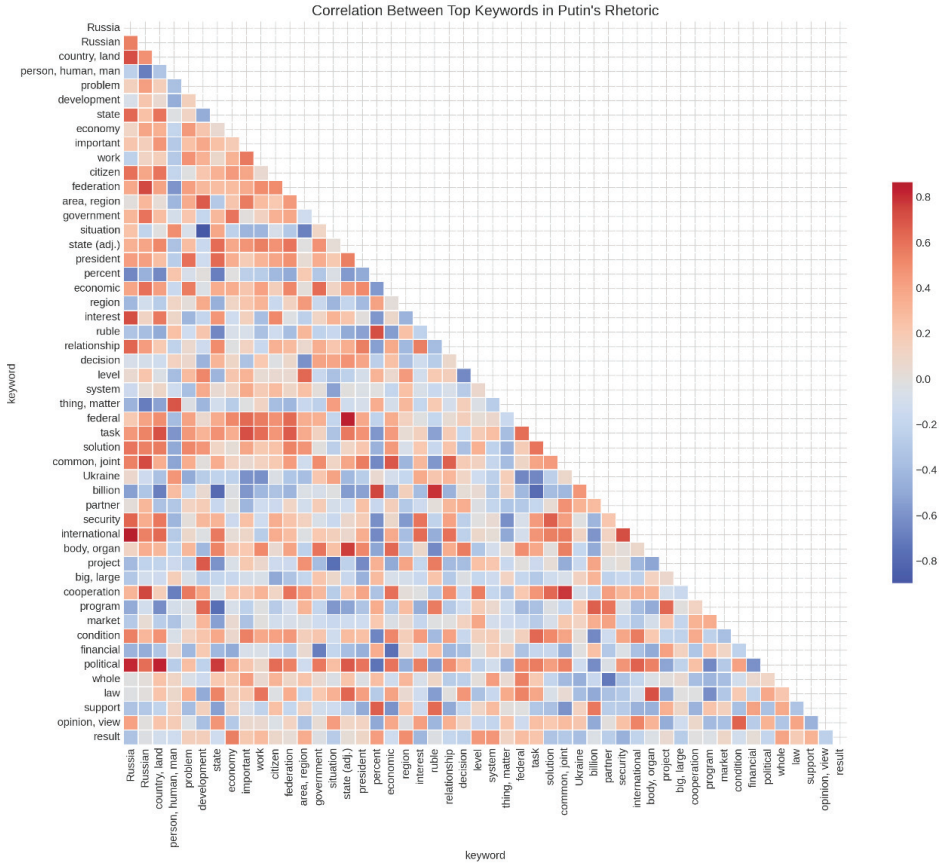

Fig. 4. Correlation between the top 50 keywords in Putin's rhetoric

This correlation heat-map reveals which keywords tend to appear together in Putin's rhetoric, helping to identify conceptual associations in his political discourse.

4.2.1 Key findings by political event

Chechen war period (1999–2000): During this early period, '*Chechnya*' was the second most prominent keyword after '*Russia*', reflecting the centrality of the Chechen conflict to Putin's early political identity. The rhetoric focused heavily on state security, terrorism, and the establishment of federal authority.

Putin's return to the presidency (2012): Keywords show a shift towards more domestic governance and economic development themes, with a continued emphasis on Russia's international position.

Annexation of Crimea (2014): After the annexation of Crimea, '*Ukraine*' emerges as a significant keyword in Putin's rhetoric.

Ukraine invasion period (2022–2024): In the final period, we see a consistent pattern of keywords with '*Russia*' remaining dominant, but with '*Ukraine*' maintaining a significant presence. The rhetoric continues to emphasise Russian identity, citizenship, and state interests.

This inspired us to investigate whether there is a correlation between keyword importance and historical events (see Fig. 5).

4.3 Trend of average keyword importance in Putin's speeches

The graph in Fig. 5 shows the temporal evolution of the average TF*IDF scores of keywords in Vladimir Putin's speeches over time.

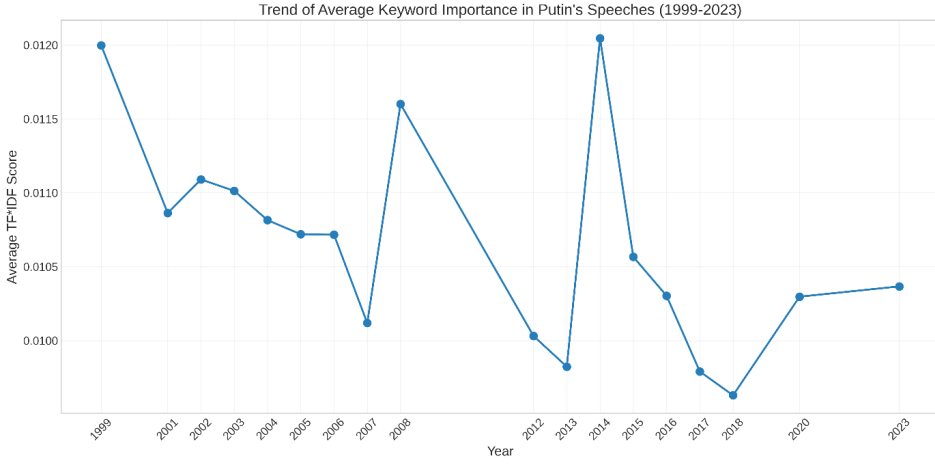

Fig. 5. Trend of average keyword importance in Putin's speeches from 1999 to 2024 according to TF*IDF score

4.3.1 Key observations

Initial peak (1999): The graph begins with a high average TF*IDF score (~0.0120) in 1999, reflecting a strong emphasis on certain keywords during Putin's early political career, likely related to his initial rise to power and the need to establish a clear and coherent rhetorical narrative when establishing nationalist themes was crucial.

Decline (2000–2006): From 2000 to 2006, there is a gradual decline in the average importance of keywords, with the score stabilising at 0.0105. This period corresponds to the first period of Putin's presidency; which may indicate a diversification of topics or a broadening of the discourse as the leadership stabilised.

Spike (2008): The sharp increase in 2008 marks a significant rhetorical shift or focus, and the average TF*IDF score peaks again. This may be related to the global financial crisis and the political transition, with Putin changing roles and emphasising a strong, unified discourse in a time of uncertainty.

Low point (2013): The lowest point in the graph occurs in 2013, with an average score below 0.0100. This period precedes the annexation of Crimea in 2014, suggesting a possible lull in focused rhetorical emphasis.

Major peak (2014): The dramatic increase in 2014 is associated with the annexation of Crimea and heightened geopolitical tensions. This suggests a concerted use of specific keywords to address the crisis and consolidate domestic and international support.

Decline and stabilization (2015–2024): After 2014, the average TF*IDF score gradually declines, suggesting a possible moderation in language intensity, reaching another low around 2018. However, there is a slight upward trend from 2020 onwards, which may reflect recent domestic or global challenges that prompt a recalibration of rhetoric, such as the COVID-19 pandemic or the invasion of Ukraine.

5 CONCLUSION

The quantitative corpus analysis of Putin's speeches over the past 25 years reveals significant stylistic and thematic shifts in his political rhetoric. Two main periods – before and after 2012 – show a clear change in rhetorical strategy and thematic focus associated with major political events, and coinciding with Putin's return to the presidency and the intensification of Russia's authoritarian governance and geopolitical assertiveness. Thematic clusters identified on the basis of 330 unique keywords confirm a continued emphasis on national identity, but shifting from internal national unity-building to internationally oriented geopolitical issues. Economic and social issues gradually gain in importance, while security and conflict issues dominate during periods marked of military crises. The temporal evolution of keyword importance, as measured by TF*IDF scores, shows notable peaks corresponding to critical moments in Russian politics, including an initial peak in 1999, a sharp increase in 2008, a dramatic increase in 2014 related to the annexation of Crimea, and subtle changes in the period 2020–2024. These findings provide a comprehensive and systematic understanding of the evolution of Putin's language, showing how his discourse adapts to changing political circumstances and strategic goals.

ACKNOWLEDGEMENTS

The research was supported by the Ministry of Education of the Czech Republic IGA_FF_2025_042 "The Czech Language and Its Worlds: Interdisciplinary Approaches to Textual Study in the 21st Century".

References

Al-Manaseer, A. (2025). Deconstructing the Ideological Frame of President Vladimir Putin's Rhetoric: A Socio-Cognitive Analysis. Wasit Journal for Human Sciences, 21(1), pp. 984–999.

Burrows, J. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing, 17(3), pp. 267–287.

Chiluwa, I., and Ruzaite, J. (2024). Analysing the language of political conflict: a study of war rhetoric of Vladimir Putin and Volodymyr Zelensky. Critical Discourse Studies, pp. 1–17.

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

Fafiyebi, D. O., and Fafiyebi, O. F. (2025). A Speech Act Analysis of the Utterances of Selected Key Actors in the Russian/Ukrainian Crisis: Pragmatics. International Journal of Language and Literary Studies, 7(1), pp. 336–352.

Hidalgo-Cobo, P., López-Marcos, C., and Puebla-Martínez, B. (2024). Discourse analysis from an international relations perspective: the case study of Tucker Carlson's televised interview with Vladimir Putin. aDResearch ESIC International Journal of Communication Research, 32 (November, 2024), e285.

Hoover, D. L. (2003). Multivariate Analysis and the Study of Style Variation. Literary and Linguistic Computing, 18(4), pp. 341–360.

Janda, L., Fidler, M., Cvrček, V., and Obukhova, A. (2022). The case for case in Putin's speeches. Russian Linguistics, 47, pp. 15–40.

Kadim, E. (2023). A Critical Discourse Analysis of Vladimir Putin's Speech Announcing 'Special Military Operation' in Ukraine. International Journal of Humanities and Educational Research, 5, pp. 424–444.

Kopik, M. (2023). Comparative analysis of American and Russian political discourse: A discourse analysis study. Linguistics Beyond and Within, 9, pp. 49–59.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). Mining of massive datasets. Accessible at: http://www.mmds.org/#ver30.

Libovický, J. (2016). KER – Keyword extractor. [software] Accessible at: http://lindat.mff.cuni.cz/services/ker/.

Oleinik, A. (2023): A comparison of two text specificity measures analyzing a heterogenous text corpus. Glottometrics, 54, pp. 1–12.

Rajaraman, A., and Ullman, J. D. (2011). Data mining. In: J. Leskovec – A. Rajaraman – J. D. Ullman (eds.): Mining of Massive Datasets, pp. 1–19. Cambridge.

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In Proceedings of the First Instructional Conference on Machine Learning, pp. 133–142. Piscataway, NJ: Rutgers University.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), pp. 513–523.

Scott, M., and Tribble, Ch. (2006). Textual patterns. Key words and corpus analysis in language education. Amsterdam: Benjamins.

Shahbaz, J., and Nawab, H. (2024). Language, Politics, and Power: Unveiling Putin's Annexation Narrative through Fairclough's Model, 7(2), pp. 24–33.

stylo: Stylometric Multivariate Analyses (version 0.7.4). [software]. Accessible at: https://cran.r-project.org/web/packages/stylo/index.html.

The R Project for Statistical Computing: R (version 4.2.0) [software]. Accessible at: https://www.r-project.org/.

Tutar, H., and Bağ, S. M. (2023). Critical discourse analysis on leader statements in the Russia-Ukraine War. Etkileşim, 11, pp. 44–66.

Wang, Y., and Zeng, T. (2023): Fellow or foe? A quantitative thematic exploration into Putin's and Trump's stylometric features. Glottometrics, 54, pp. 39–57.