Language Complexity in Populist Rhetoric

Sergio E. Zanotto[1], **Diego Frassinelli**[1,2], **Miriam Butt**[1]
[1]Department of Linguistics & Cluster of Excellence “The Politics of Inequality”, University of Konstanz
[2]Center for Information and Language Processing, LMU Munich
{sergio.zanotto, diego.frassinelli, miriam.butt}@uni-konstanz.de

Abstract

Research suggests that politicians labeled as populists tend to use simpler language than their mainstream opponents. Yet, the metrics traditionally employed to assess the complexity of their language do not show consistent and generalizable results across different datasets and languages. This inconsistencies raise questions about the claimed simplicity of populist discourse, suggesting that the issue may be more nuanced than it initially seemed. To address this topic, we analyze the linguistic profile of IMPAQTS, a dataset of transcribed Italian political speeches, to identify linguistic features differentiating populist and non-populist parties. Our methodology ensures comparability of political texts and combines various statistical analyses to reliably identify key linguistic characteristics to test our case study. Results show that the “simplistic” language features previously described in the literature are *not* robust predictors of populism. This suggests that the characteristics defining populist statements are highly dependent on the specific dataset and the language being analysed, thus limiting the conclusions drawn in previous research. In our study, various linguistic features statistically differentiate between populist and mainstream parties, indicating that populists tend to employ specific well-known rhetorical strategies more frequently; however, none of them strongly indicate that populist parties use simpler language.

1 Introduction

The concept of populism has gained a huge focus in social sciences, with different scholars attempting to systematically analyse the phenomenon to understand its core components (e.g., Huguet Cabot et al., 2021; Pérez-Curiel et al., 2021; Klamm et al., 2023). For instance, inspired by the *social identity theory* of Tajfel and Turner (2004), different studies have employed Natural Language Processing (NLP) techniques to studying social group appeals[1] in political texts (Huber, 2022; Licht and Sczepanski, 2023; Zanotto et al., 2024). Several studies have explored the rhetorical power of identity appeals to citizens and their effects on voting behavior (e.g., Strom, 1990; Wodak, 2012; Thau, 2019). This rhetorical power is evident especially when focusing on populist communication, where the tendency to appeal to “the people” is considered a universal component of all different realizations of populism (Canovan, 2004; Laclau, 2006). Populist parties divide society in two groups: "pure people" and "corrupt elite" and advocate for politics to represent the general will of the people (Mudde, 2004; Jagers and Walgrave, 2007). Therefore, language complexity becomes both an important characteristic of populist communication, as well as a tool for appealing to a broader public of ordinary people (Decadri and Boussalis, 2020; McDonnell and Ondelli, 2022). This assumption lies on the idea that "the people" are less-educated and therefore speak simpler. Simple language helps citizens to better understand political positions (Senninger, 2023), and scholars claim populists use it to convey their simplistic message and strengthen their positions as part of "the people" (Canovan, 1999; Zaslove, 2008; McDonnell and Ondelli, 2022). Thus, to describe populist language, different researchers have analysed political texts throughout syntactic and lexical features like readability scores, type-token ratio analysis, and dictionary approaches (Rooduijn and Pauwels, 2011; Bischof and Senninger, 2018), showing that populist parties generally employ simpler language than their mainstream opponents. However, different studies have highlighted very diverse patterns in the language of populism, questioning the validity of using lan-

[1] Social group appeals refer to strategies of communication that target specific groups based on shared characteristics, such as ethnicity, religion, socioeconomic status, or political affiliation (Huber, 2022).

guage complexity as a distinguishing feature for populism. (Trotta et al., 2019; McDonnell and Ondelli, 2022).

In this paper, we investigate what are the distinctive features that set populist from non-populist parties apart. Compared to existing studies, our analysis focuses on speeches within the Italian political arena, extracted from the IMPAQTS corpus (Cominetti et al., 2022). We categorized the discourses in IMPAQTS as either populist or non-populist given the political affiliations of the speakers, as outlined in previous research such as Di Cocco and Monechi (2022). The categorization of populist and non-populist parties rely on the classification from "The PopuList 3.0" (Rooduijn et al., 2023).

Our main contributions are: (i) we challenge the prevailing notion that populism is characterized by simpler language; and (ii) we identify specific linguistic features that indicate a tendency in using well-known rhetorical strategies; (iii) we propose a systematic approach to empirically select linguistic features that differentiate populist and non-populist discourses in our dataset.

2 Related Work

Scholars investigate how populist politicians influence the public opinion via their discourses (Canovan, 2004; Laclau, 2006). Among the different definitions of *populism*, the division of society in two groups, namely the "pure people" and the "corrupt elite", is considered a universal feature of all populist parties (Mudde, 2004; Jagers and Walgrave, 2007). This is the definition we adopt in our research. Several studies measure populism in text by looking at *how* and *to whom* populists refer in their discourses (Jagers and Walgrave, 2007; Huguet Cabot et al., 2021; Klamm et al., 2023). In this way, they show how the indexing of people and the anti-establishment rhetoric are typical characteristics of populist communication. For instance, Rooduijn and Pauwels (2011) and Decadri and Boussalis (2020) conduct a semi-automatic content analysis using dictionaries of words related to populist rhetoric, such as *citizen, people, caste, elite*.

2.1 Language complexity

Even though the use of simple and accessible language is considered a tool for appealing to a broader audience of "ordinary people" (Decadri and Boussalis, 2020; McDonnell and Ondelli, 2022), to date, there is no agreement on which computational measures best describe how complex a language is (Ehret et al., 2021). The literature indicates that evaluating language complexity requires to analyze both syntactic and lexical information (Ehret et al., 2023). Consequently, focusing solely on textual complexity, often measured through readability scores, captures only one facet of it. Of the various definitions of language complexity available in the literature (Pallotti, 2015), we adopt the one from second language acquisition (SLA), especially the definition of *structural complexity* as “a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns” (Pallotti, 2015). Therefore, in order to estimate the complexity of a text, it is necessary to analyze it through its linguistic dimensions.

2.2 Complexity of political texts

Many studies rely on readability scores to assess the complexity of political texts (e.g., Spirling (2016); Bischof and Senninger (2018); Schoonvelde et al. (2019); Decadri and Boussalis (2020); Senninger (2023)). Readability scores are language specific and assess textual complexity by analyzing elements such as the number of words, sentences, and characters. For example, the Flesch-Kincaid readability tests (Kincaid et al., 1975) are tailored for English, using sentence length and syllable count. The Gulpease Index (Lucisano and Piemontese, 1988), used for Italian texts, considers the number of characters per word and words per sentence.

Given the criticisms regarding the validity of readability scores for measuring text complexity (see Chall (1996) for an extensive overview of these criticisms), alternative measures have been employed to quantify the textual complexity and the syntactic complexity of political discourses. These measures include the number of tokens in a document, as well as the length of its words, its sentences and its syntactic complexity (Tolochko and Boomgaarden, 2019; McDonnell and Ondelli, 2022). Syntactic complexity is typically analysed through syntactic depth or syntactic dependency (Tolochko and Boomgaarden, 2019). Syntactic depth considers the number of nested clauses or phrases within each sentence, while syntactic dependency measures the distance between a syntactic head and its farthest dependent for each sentence. They are used as a better fitting measures of language complexity for spoken language.

Another level of analysis pertains to lexical complexity, and it involves the use of type-token ratio, lexical density and the use of frequent words (e.g., Schoonvelde et al. (2019); Trotta et al. (2019); Takikawa and Sakamoto (2020); McDonnell and Ondelli (2022)). However, according to the literature, these features are not always significant across different studies in distinguishing populist and non-populist discourses (Trotta et al., 2019; McDonnell and Ondelli, 2022).

3 Data

In this section, we describe the dataset used for our analysis and we illustrate the criteria of classification of populist and non-populist parties.

Dataset

We use the IMPAQTS corpus (Cominetti et al., 2022) to identify linguistic features that distinguish between populist and non-populist parties. IMPAQTS is a corpus containing circa 1,500 transcripts of Italian political speeches from 1948 to 2023. We select this corpus as it is the biggest corpus available of multi-genre speeches of Italian politicians. The nature of these discourses is monological. There are six different genres of speech, namely rallies, parliamentary speeches, party meetings, face-to-face declarations, transmitted declarations, and new media declarations. We restrict our analysis to discourses from 1994 onwards, aligning with the emergence of the first populist parties in Italy (e.g., *Forza Italia*, "Forward Italy"). We further filter the data by keeping only 88 politicians having at least eight documents each. Thus, in our analyses we include 851 documents, 369 (43%) of which are labelled as populist. Table 2 in the Appendix reports the number of documents and tokens for each politician included in our analysis.

Classification of populist parties

We rely on "The PopuList 3.0" (Rooduijn et al., 2023) to extract the list of populist and non-populist parties for our feature analysis. The definition used to classify parties as populist relies on the Mudde (2004) identification of "The People" vs. "The Elitè" distinction and their view of politics as expression of the general will of the people. The classification of parties in "The PopuList 3.0" (Rooduijn et al., 2023) was conducted using an 'Expert-informed Qualitative Comparative Classification' (EiQCC). This method uses experts of political communication who qualitatively compare and classify political parties based on their expertise. Table 3 in the Appendix provides a list of Italian populist parties.

4 Methodology

Analyzing language involves dealing with several challenges, like the need for selecting among a vast number of features and the strong collinearity between different language features. In this section, we present the main features used in our analysis and the control features used to guarantee the comparability of the different texts. Then, we illustrate the feature selection procedure and the logistic regression models used to assess the statistical significance of the selected features.

4.1 Features Collection

In total, we collected 147 features from different linguistic levels of analysis. All features are included in the selection mechanism.

4.1.1 Features derived from the literature

In our analysis, we include the six features mostly used in the literature to quantify language complexity in populist and non-populist parties.

Raw text Parameters

Gulpease Index: `G_index` (Lucisano and Piemontese, 1988) is the Italian measure for readability in text. This measure suggests that the higher the level of readability, the easier the text is.

Characters per token: `Char_per_Tok` are calculated with the "UD profiling" tool (Brunato et al., 2020) and represents the average length of words per document. The interpretation of this measure suggests that the longer the words in a text, the more complex the text is.

Lexical Features

Lexical density: `Lexical_density` is calculated using the "UD profiling" tool, and it consists in the number of content words divided by the total number of words. This measure indicates the degree of use of content words in a text, suggesting that the higher the degree, the more informative the text.

Type-token ratio: `Type_token_ratio` is calculated by counting the number of unique tokens and divide it by the total number of tokens. We include this feature to verify if populist texts tend to have a lower lexical diversity compared to non-populist texts.

Word frequency: Word_frequency is calculated using a frequency list[2] and, based on the way we calculated it, indicates that the greater the score, the less frequent words are used in a text.

Syntactic measures
Syntactic depth: the average maximal depth (Avg_max_depth) is calculated using the "UD profiling" tool. The intepretation of this measure indicates that the greater the average depth of syntactic trees in sentences, the more complex the text is.

4.1.2 Other tested features

We extend our feature analysis by using the "UD profiling" tool for profiling the linguistic style of each text. Moreover, we include "Age of Acquisition" and "Concreteness" as plausible features in differentiating populist and non-populist rhetoric. Finally, we add a measures of people-centric and anti-elitè rhetoric as in Decadri and Boussalis (2020).

UD Profiling's features: UD Profiling's features are 141 features measured using the "UD profiling" tool. They can be grouped as follows: Raw Text Properties, Lexical Variety, Morphosyntactic Information, Verbal Predicate Structure, Global and Local Parse Tree Structures, Syntactic Relations, and Use of Subordination. A detailed list of the UD profiling's features can be found in Table 4 in the Appendix.
Age of Acquisition: AoA is calculated using the vocabulary in Montefinese et al. (2019). This parameter is calculated summing the age of acquisition of each word in the text and dividing it by the total number of tokens in the text. We include this feature to verify if populist texts tend to use simpler, earlier acquired words compared to non-populist texts.
Concreteness: Concreteness is calculated using the vocabulary in Gregori et al. (2020). This parameter is calculated summing the concreteness score of each word in the text and dividing it by the total number of tokens. We include this feature to verify if populist texts tend to use more concrete, tangible words compared to non-populist texts.

People-centric and anti-elitè rhetoric
Populist words ratio: The ratio of using populist words (Populist_words_ratio) is calculated using the dictionary approach in Decadri and Boussalis (2020), without distinguishing anti-elitism and people-centric rhetoric.[3] Table 5 in Appendix shows the seed words of the dictionary. We include this feature to verify if populists tend to use more people-centric and anti-elitè words compared to non-populists.

4.2 Control Features

We focus on the comparability of political texts and their metadata to guarantee a reliable analysis of their linguistic components. By using control features in our regression analyses, we account for potential confounding variables, thereby enhancing the accuracy and comparability of our modeling study. For each political text, it is fundamental to control for the following metadata extracted from the IMPAQTS corpus:
Time: Decade includes span of 10 years from the 1994 until 2023.
Genre: Type consists of 6 different genres of transcribed speeches. The institutional setting varies among the speeches (e.g. Rallies vs Parliamentary speeches), making them clearly different from a theoretical perspective.
Author: Author refers to the politician that acts as the speaker of the speech.
Topic(s): Topic is the main argument of one document. We apply a Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to identify the dominant topic of each text (see Table 9 in the Appendix for further details).
Author's role: Is_Majority refers to the government/opposition role of the speaker's party during the date of the document.
Author's political party: Political_Party refers to the affiliation's party of the author at the date of the speech.
Transcriber: Transcriber refers to the person who transcribed the speech. It does not apply to written texts.

[2] https://invokeit.wordpress.com/frequency-word-lists

[3] Populist rhetoric is typically divided into two components: anti-elitism and people-centric rhetoric (refer to Section 2). We aggregate these components to focus on the general level of populist rhetoric.

4.3 Study: Populism Classification

We focus on the multifaceted concept of populism as a case study for profiling political texts and verifying different communication strategies, given a politician affiliation with populist parties. In our analysis, the classification of a document as populist relies on the author's political party. A score of 1 is given to parties classified as populist, 0 viceversa. For details on the classification of parties as populists see Section 3.

We streamlined a methodological framework that enhances the reliability of linguistic profile analyses within political texts. All codes are accessible at `https://github.com/Sergio-E-Zanotto/language_complexity_populism`.

4.3.1 Data Pre-Processing

To obtain a balanced corpus, we selected 88 authors represented by at least eight texts (refer to Section 3). Given that different features come with very different scales, we pre-process our data by standardizing all the numerical variables.

4.3.2 Feature Selection

We apply LASSO regression (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) to automatically identify the most relevant linguistic measures among all our 147 features. LASSO is a logistic regression method that includes a penalty term, which is the absolute value of the magnitude of the coefficients. This penalty term encourages the reduction of less important feature coefficients to zero, thereby performing feature selection and regularization to enhance the prediction accuracy and interpretability of the model. We automatically scored the penalty term λ (0.199) to address collinearity issues through our feature selection process. After each logistic regression, we apply the Variance Inflation Factor (VIF) to ensure that no collinearity remains.

4.3.3 Features Analysis

In our analysis, we utilize logistic regression to identify the statistically significant features that differentiate populist and non-populist discourses. First, we test only the features derived from the literature to assess their importance in distinguishing populists and non-populists discourses (see Section 4.1.1). Subsequently, we consider all features for our analysis and we select the top 15 predictors[4] that the feature selection process indicated as the most important in distinguishing between populist and non-populist discourses (refer to Section 4.3.2). First, we analyse the features with a logistic model to verify differences among populist and non-populist parties, without accounting for the communication style of each individual politician or any possible effect of the process of transcriptions. Second, we utilize a general mixed-effects model to add author and transcriber effects as random structure. All the regressions include control features (see Section 4.2).

5 Results

5.1 Features analysis on Populism

Table 1 reports the mean value of each linguistic feature for populist and non-populists parties and their difference (populist−non-populist). It also indicates which predictors reach significance according to the logistic regression (GLM) and the mixed-effects logistic regression (GLMER) models. Respectively, Table 6 and Table 7 in the Appendix report all the details of our statistical analyses.

According to the GLM model, `Lexical_Density` is the only feature derived from the literature that is significant in classifying populism, and it shows how populists utilize a slightly higher number of content words. We can appreciate from our selection of features how the degree of proper nouns (`Upos_dist_PROPN`) is significantly higher in populist texts. Additionally, populist texts show a higher ratio of populist words (`Populist_words_ratio`) and a higher number of second-person singular verbs (`Verbs_num_pers_dist_Sing2`). Furthermore, the percentage of verbal roots (`Verbal_root_perc`) is slightly lower in populist texts. The distribution of determiners and predeterminers (`Dep_dist_det_predet`) is also notably higher in populist texts. In Italian, this relation is used for the lemmas *tutto* ('all'), *entrambi* ('both'), and *ambedue* ('both'), when they appear in front of another determiner. We can also see that the degree of adjectives (`Upos_dist_ADJ`) is significantly lower in populist texts.

[4] We selected the top 15 features, which represent approximately 10% of the total features, to focus on the most impactful predictors while maintaining a manageable number of variables for the analysis. The features that are not significant are not reported in the paper.

Predictor	Populist	Non-Populist	Difference	Significance	
				GLM	GLMER
G_index	52.063	52.037	0.026		
Char_per_tok	4.703	4.705	-0.002		
Type_token_ratio	0.406	0.408	-0.002		
Word_frequency	0.596	0.602	0.006		
Avg_max_depth	5.629	5.812	-0.183		
Lexical_density	0.471	0.469	0.002	**	
Upos_dist_PROPN	2.744	2.237	0.507	***	
Dep_dist_det_predet	0.225	0.180	0.045	***	
Populist_words_ratio	0.008	0.007	0.001	***	
Verbs_num_pers_dist_Sing2	2.066	1.584	0.482	**	
Verbal_root_perc	84.680	86.418	-1.738	*	
Verbs_mood_dist_Cnd	1.177	1.487	-0.310	*	**
Verbs_form_dist_Fin	43.163	45.105	-1.942	***	**
Upos_dist_ADJ	5.241	5.580	-0.339	***	
Subordinate_dist_4	1.255	0.951	0.304	*	*
Verb_edges_dist_1	14.129	13.508	0.621	*	**

Table 1: Comparison of linguistic predictors between *Populist* and *Non-Populist* groups along with their differences (Populist-Non-Populist). Statistical significance * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The GLMER model that includes author and transcriber effects shows that most of those features lose significance, and neither features derived from the literature nor populist words remains robust predictors of populism. We can attribute such change to the high variance in the author group (see Table 7 in the Appendix). However, a few features remain robust and indicate distinct patterns in populist language. Specifically, populists use fewer conditionals (`Verbs_mood_dist_Cnd`) and fewer finite verbs (`Verbs_form_dist_Fin`) than non-populists. Additionally, they tend to use more verbs with valency 1 (`Verb_edges_dist_1`) and employ subordinate clauses in chains of four (`Subordinate_dist_4`).

Moreover, when comparing models with random effects, the model informed with our automatically selected features performs better in terms of AIC (Akaike Information Criterion) than the one informed by the features derived from the literature (see Table 8 in the Appendix).

6 Discussion

Our analysis of multiple linguistic features yields several insights. First, traditional language complexity features identified in populism research do not robustly transfer to our data, often failing to distinguish populist discourse effectively. This suggests that the characteristics defining populist statements are highly dependent on the specific dataset and the language analysed, thus limiting the general conclusions drawn in previous research. Second, our feature selection revealed interesting trends when comparing populist vs. non-populist parties, particularly the well-known difference in the use of populist words. According to the literature, populists often emphasize a dichotomy between "the people" and "the elite" to rally support (Mudde, 2004). The most significant features also indicate a much wider use of proper nouns and quantifiers such as "all" and "both" as predeterminers by populist parties. This could imply a tendency to make absolute statements and to generalize broadly, reinforcing the populist narrative of representing the entire population against a unified elite, as exemplified in our corpus by sentence (1).

(1) [...] *perché non pensate a tutti gli italiani, pensate solo ad alcuni di essi* [...]
‘[...] because you don’t think about all Italians, you only think about some of them[...]’

Moreover, lexical density is significant in showing that populists use more content words. However, while populists use fewer adjectives, they tend to use proper nouns and second-person singular verbs more consistently. This might suggest a focus on specific individuals or groups and direct engagement with the audience, respectively. Additionally, populists use fewer verbal roots to structure their sentences, potentially indicating a reliance on more direct and straightforward statements.

When controlling for authors' effects, all these features lose significance, indicating huge variance in politicians' communication styles. Only four features remain significant after accounting for authors' effects. In IMPAQTS, we observe a trend among populist parties to use conditional verbs less frequently, potentially indicating a preference for stronger epistemic modality. For example, in our corpus, non-populists might prefer statements that convey less epistemic strength, as exemplified by sentence (2), which clearly conveys less certainty compared to a straightforward statement like 'we want to say something'.

(2) *E vorremmo, vorremmo poter dire una cosa:* [...]
'an we would, we would want to say something: [...]'

We also observe that populists use fewer finite verbs, implying a greater use of non-finite verbal forms. We notice a consistent use of nominalizations with non-finite verbs as the syntactic head of noun phrases, as exemplified by sentence (3). In adult speech, nominalization facilitates abstractness, which creates a sense of detachment and allows events to be presented as undeniable facts (Bello, 2016).

(3) *So bene che conoscere la regola dell'ascolto e del rispetto in democrazia non è cosa condivisa da tutti.*
'I know well that knowing the rule of respect in democracy is not something shared by everyone.'

Furthermore, nominalizations can be seen as a form of valency reduction in the formation of predicates (Mackenzie, 1985). We observe that populists tend to employ more verbs with a valency of 1, meaning verbs with only a single dependency link, either with an argument or a modifier. This strengthens the interpretation that populists seek to present events as undeniable, as exemplified in the corpus by sentence (4).

(4) [...] *la gente vuole tornare a contare,* [...] *a contare, a decidere, accogliere chi vuole accogliere, espellere chi vuole espellere* [...]
'[...] people want to matter again, [...] to matter, to decide, to welcome those they want to welcome, to expel those they want to expel [...]'

Finally, the use of subordinate clauses in chains of four shows a tendency for populists to employ repetitions in their sentences, as in sentence (5). This technique emphasizes the key points and creates a memorable rhythm, akin to the well-known rhetorical strategy known as the "rule of three" (Barry, 2018).

(5) *perché voi lo sapete, io credo nel consiglio comunale, credo nei dibattiti consiliari, credo che questo sia un fulcro forte della democrazia.*
'because you know, I believe in the city council, I believe in council debates, I believe that this is a strong cornerstone of democracy.'

Overall, our models do not strongly suggest that populist parties use simpler language than their mainstream rivals. We argue that substantial differences can be found in the simplicity of the conveyed content, more than in the simplicity of the language used to convey it, as discussed in McDonnell and Ondelli (2022). Instead, our results suggest that populists adhere more to specific, well-known rhetorical strategies, making populism a communication strategy that is common to very diverse parties and politicians. Indeed, in our corpus, sentence (6) is the perfect example of a combination of the above characteristics. The use of copular "be" conveys a stronger epistemic modality and affirms the undeniability of the stated facts, while the repetitions in the sentence help to emphasize key points and create a memorable rhythm.

(6) *La crisi non c'è, la crisi non esiste, c'è il pessimismo e non date retta al pessimismo.*
'there is no crisis, the crisis does not exist, there is pessimism and do not listen to pessimism.'

7 Conclusion and Future Work

In our analysis of the linguistic characteristics of Italian political speeches, we implemented a detailed methodology to ensure the comparability of texts and utilized a feature selection process to explore linguistic differences among populists and non-populists parties. Our study reveals that traditionally employed features of language complexity derived from the literature do not show statistical significance in distinguish populist and non-populist discourse in the IMPAQTS corpus. This inconsistency underscores the importance of context and corpus specificity in linguistic analyses, cautioning against overgeneralizing findings.

Moreover, while we observed an increased occurrence of populist rhetoric —characterized by themes of people-centrism and anti-elitism— in speeches from aggregating populist parties, this did not coincide with simpler language use. Especially, most of these features were not robust to the individuality of speakers communication style within our dataset. We highlight the tendency of populists' speaker to employ specific, well-known rhetorical strategies in their speeches. However, our research highlights again the need for nuanced analysis that considers the diverse characteristics of the corpus being studied.

Building on this foundation, future research will aim to enhance the granularity of populism annotation in textual data, following approaches like those outlined by Klamm et al. (2023). Additionally, examining other features of political communication, such as emotional content as suggested by Huguet Cabot et al. (2021), may offer deeper insights into the nuances of populist rhetoric across different authors and political parties. This direction promises to refine our understanding of the linguistic strategies employed within political discourse.

8 Limitations

One limitation of our study involves the nature of the corpus analyzed. The controls within IMPAQTS present challenges due to their unbalanced nature, making it difficult to aggregate the results. For example, this imbalance may potentially favor more frequent genres, such as parliamentary speeches, over smaller ones. Despite this, the significance of incorporating controls to enhance the robustness of our findings remains undisputed.

References

Patrick Barry. 2018. The rule of three. *Legal Communications and Rhetoric: JALWD*, 15:247.

Iria Bello. 2016. Cognitive implications of nominalizations in the advancement of scientific discourse. *International Journal of English Studies*, 16(2):1–23.

Daniel Bischof and Roman Senninger. 2018. Simple politics for the people? Complexity in campaign messages and political knowledge. *European Journal of Political Research*, 57(2):473–495.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France.

Margaret Canovan. 1999. Trust the people! Populism and the two faces of democracy. *Political Studies*, 47(1):2–16.

Margaret Canovan. 2004. Populism for political theorists? *Journal of Political ideologies*, 9(3):241–252.

Jeanne S Chall. 1996. Varying approaches to readability measurement. *Revue qué Bécoise de linguistique*, 25(1):23–40.

Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, and Alessandro Panunzi. 2022. Impaqts: un corpus di discorsi politici italiani annotato per gli impliciti linguistici. In *Corpora e Studi linguistici. Atti del LIV Congresso della Società di Linguistica Italiana (Online, 8–10 settembre 2021), a cura di Emanuela Cresti e Massimo Moneglia. Milano, Officinaventuno*, pages 151–164.

Silvia Decadri and Constantine Boussalis. 2020. Populism, party membership, and language complexity in the italian chamber of deputies. *Journal of Elections, Public Opinion and Parties*, 30(4):484–503.

Jessica Di Cocco and Bernardo Monechi. 2022. How populist are parties? Measuring degrees of populism in party manifestos using supervised machine learning. *Political Analysis*, 30(3):311–327.

Katharina Ehret, Aleksandrs Berdicevskis, Christian Bentz, and Alice Blumenthal-Dramé. 2023. Measuring language complexity: challenges and opportunities. *Linguistics Vanguard*, 9(s1):1–8.

Katharina Ehret, Alice Blumenthal-Dramé, Christian Bentz, and Aleksandrs Berdicevskis. 2021. Meaning and measures: Interpreting and evaluating complexity metrics. *Frontiers in Communication*, 6:640510.

Lorenzo Gregori, Maria Montefinese, Daniele P Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. CONCRETEXT@EVALITA2020: The concreteness in context task. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR.org.

Lena Maria Huber. 2022. Beyond policy: the use of social group appeals in party communication. *Political Communication*, 39(3):293–310.

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. them: A dataset of populist attitudes, news bias and emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945, Online. Association for Computational Linguistics.

Jan Jagers and Stefaan Walgrave. 2007. Populism as political communication style: An empirical study of political parties' discourse in Belgium. *European Journal of Political Research*, 46(3):319–345.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Tech. Rep.*

Christopher Klamm, Ines Rehbein, and Simone Paolo Ponzetto. 2023. Our kind of people? Detecting populist references in political debates. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1227–1243, Dubrovnik, Croatia. Association for Computational Linguistics.

Ernesto Laclau. 2006. On populist reason. *Tijdschrift Voor Filosofie*, 68(4):832–835.

Hauke Licht and Ronja Sczepanski. 2023. Who are they talking about? detecting mentions of social groups in political texts with supervised learning. OSF Preprints, 20 June 2023.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana. *Scuola e Città*, pages 110–124.

J Lachlan Mackenzie. 1985. Nominalization and valency reduction. *Predicates and Terms in Functional Grammar. Dordrecht: Foris*, pages 31–51.

Duncan McDonnell and Stefano Ondelli. 2022. The language of right-wing populist leaders: Not so simple. *Perspectives on Politics*, 20(3):828–841.

Maria Montefinese, David Vinson, Gabriella Vigliocco, and Ettore Ambrosini. 2019. Italian age of acquisition norms for a large set of words (itaoa). *Frontiers in Psychology*, 10:278.

Cas Mudde. 2004. The populist zeitgeist. *Government and Opposition*, 39(4):541–563.

Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Concha Pérez-Curiel, Rubén Rivas-de Roca, and Mar García-Gordillo. 2021. Impact of Trump's digital rhetoric on the us elections: A view from worldwide far-right populism. *Social Sciences*, 10(5):152.

Matthijs Rooduijn and Teun Pauwels. 2011. Measuring populism: Comparing two methods of content analysis. *West European Politics*, 34(6):1272–1283.

Matthijs Rooduijn, Andrea LP Pirro, Daphne Halikiopoulou, Caterina Froio, Stijn Van Kessel, Sarah L De Lange, Cas Mudde, and Paul Taggart. 2023. The populist: A database of populist, far-left, and far-right parties using expert-informed qualitative comparative classification (eiqcc). *British Journal of Political Science*, pages 1–10.

Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N Bakker. 2019. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PloS One*, 14(2):e0208450.

Roman Senninger. 2023. What makes policy complex? *Political Science Research and Methods*, 11(4):913–920.

Arthur Spirling. 2016. Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832–1915. *The Journal of Politics*, 78(1):120–136.

Kaare Strom. 1990. A behavioral theory of competitive political parties. *American journal of political science*, pages 565–598.

Henri Tajfel and John C. Turner. 2004. The social identity theory of intergroup behavior. In *Political Psychology*, pages 276–293. Psychology Press.

Hiroki Takikawa and Takuto Sakamoto. 2020. The moral–emotional foundations of political discourse: a comparative analysis of the speech records of the us and the japanese legislatures. *Quality & Quantity*, 54:547–566.

Mads Thau. 2019. How political parties use group-based appeals: Evidence from britain 1964–2015. *Political Studies*, 67(1):63–82.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Petro Tolochko and Hajo G Boomgaarden. 2019. Determining political text complexity: Conceptualizations, measurements, and application. *International Journal of Communication*, 13:21.

Daniela Trotta, Sara Tonelli, Alessio Palmero Aprosio, and Elia Annibale. 2019. Annotation and analysis of the polimodal corpus of political interviews. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*.

Ruth Wodak. 2012. Language, power and identity. *Language Teaching*, 45(2):215–233.

Sergio E. Zanotto, Qi Yu, Miriam Butt, and Diego Frassinelli. 2024. GRIT: A dataset of group reference recognition in Italian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7963–7970, Torino, Italia. ELRA and ICCL.

Andrej Zaslove. 2008. Here to stay? Populism as a new party type. *European Review*, 16(3):319–336.

A Appendix

Table 2 presents the list of politicians analyzed with the number of documents and tokens available (see Section 3); Table 3 shows the list of populist parties used for our classification (see Section 3); Table 4 reports the list of all features extracted using the profiling UD's tool (Brunato et al., 2020); Table 5 provides the list of seed words of the dictionary in Decadri and Boussalis (2020) used to calculate the rate of populist words in each discourse (see Section 4.1.2).

A.1 Statistical Model Details

Furthermore, we provide all the details about the logistic regression analyses as presented in Section 5.1. Tables 6 presents the logistic regression (GLM) analysis on the most used features from the literature for analyzing language complexity in political texts and the significant features extracted by our feature selection procedure (see Section 4.3.2). Table 7 presents the mixed-effects logistic regression model (GLMER), accounting for possible author effects and transcriber effects. Controls are present in all regressions (see 4.2 for the detailed list of controls). Subsequently, Table 8 presents the comparison between mixed-effects models for the predictors derived from the literature and the automatically selected predictors (see Section 5).

A.2 Topic Analysis

In our analysis, we categorized each document based on its most prominent topic. To capture changes over time, we calculated topics at 10-year intervals. We score the optimal number of topics that better represents documents for each decade with the coherence model from Gensim python library[5]. The optimal number of topics per decade are: {'1990-1999': 3, '2000-2009': 8, '2010-2019': 7, '2020-2023': 9}. We employed Latent Dirichlet Allocation (LDA) to identify the most relevant topics for each decade, defined by the three most relevant key terms associated with each topic. Table 9 presents the topics identified for each decade, along with their corresponding key terms.

[5] https://radimrehurek.com/gensim/models/coherencemodel.html

Author	Documents	Tokens
Luigi di Maio	11	24630
Alessandra Mussolini	10	8195
Alessandro Di Battista	10	19524
Alfonso Bonafede	10	12696
Andrea Orlando	10	17292
Angelino Alfano	10	15026
Anna Finocchiaro	10	13005
Antonio di Pietro	10	12846
Beppe Sala	10	11263
Carlo Calenda	10	15466
Claudio Scajola	10	12120
Daniele Capezzone	10	13179
Danilo Toninelli	10	15469
Dario Franceschini	10	15720
Debora Serracchiani	10	15332
Enrico Letta	10	15742
Eugenia Maria Roccella	10	10600
Fabrizio Cicchitto	10	10462
Francesco Storace	10	13034
Graziano Delrio	10	14478
Guglielmo Epifani	10	16558
Ignazio La Russa	10	15546
Ignazio Marino	10	12134
Irene Pivetti	10	11958
Italo Bocchino	10	14805
Laura Boldrini	10	18068
Letizia Moratti	10	12979
Luca Zaia	10	19435
Lucia Borgonzoni	10	13854
Luigi De Magistris	10	16978
Mara Carfagna	10	12016
Maria Elena Boschi	10	16155
Maria E. Alberti Casellati	10	9563
Mariastella Gelmini	10	14552
Matteo Orfini	10	15194
Maurizio Gasparri	10	15211
Maurizio Lupi	10	15725
Monica Cirinnà	10	11682
Nichi Vendola	10	12965
Nicola Fratoianni	10	12166
Nicola Zingaretti	10	15390
Oliviero Diliberto	10	13556
Paola Binetti	10	11190
Paola Taverna	10	13409
Paolo Ferrero	10	13692
Paolo Gentiloni	10	16652
Pier Luigi Bersani	10	15914
Pietro Grasso	10	15898
Renata Polverini	10	12067
Renato Brunetta	10	12338
Renato Schifani	10	14527
Roberta Lombardi	10	12941
Roberto Calderoli	10	17480
Roberto Castelli	10	12849
Roberto Fico	10	11659
Roberto Speranza	10	13798
Rocco Buttiglione	10	11621
Romano Prodi	10	15343
Rosy Bindi	10	9892
Sandro Bondi	10	12449
Sergio Cofferati	10	9920
Stefania Prestigiacomo	10	8810
Vincenzo De Luca	10	15305
Virginia Raggi	10	15337
Walter Veltroni	10	18349
Clemente Mastella	9	16888
Daniela Santanchè	9	11394
Fausto Bertinotti	9	14338
Giorgia Meloni	9	18375
Lamberto Dini	9	11351
Mario Monti	9	12139
Matteo Renzi	9	19465
Matteo Salvini	9	22557
Maurizio Martina	9	12353
Mirko Tremaglia	9	8600
Roberto Maroni	9	11438
Silvio Berlusconi	9	17817
Anna Maria Bernini	8	10345
Antonio Tajani	8	10158
Carlo Azeglio Ciampi	8	11081
Giuseppe Conte	8	14442
Leoluca Orlando	8	9383
Sebastiano Musumeci	8	11091

Table 2: Number of documents and tokens per author in the dataset.

Political Party	Abbreviation
Lega (Nord)	LN
Forza Italia*	FI
Fratelli d'Italia	FdI
Movimento 5 Stelle	M5S
Il Popolo della Libertà	PdL
De Luca Sindaco d'Italia	DLSI
La Rete	LR
Lista Di Pietro - Italia dei Valori	IdV

Table 3: Italian populist parties - * borderline case.

Variable	Characteristics
Family: Raw Text Properties	
n_sentences	Total number of sentences
n_tokens	Total number of tokens
tokens_per_sent	Average length of sentences in a document, calculated in terms of the number of words per sentence
char_per_tok	Average number of characters per word (excluding punctuation)
Family: Lexical Variety	
ttr_lemma_chunks_100	Type/Token Ratio (TTR) calculated with respect to the lemmata in the first 100 tokens of a document. It ranges between 1 (high lexical variety) and 0 (low lexical variety)
ttr_lemma_chunks_200	Type/Token Ratio (TTR) calculated with respect to the lemmata in the first 200 tokens of a document. It ranges between 1 (high lexical variety) and 0 (low lexical variety)
ttr_form_chunks_100	Type/Token Ratio (TTR) calculated with respect to the word forms in the first 100 tokens of a document. It ranges between 1 (high lexical variety) and 0 (low lexical variety)
ttr_form_chunks_200	Type/Token Ratio (TTR) calculated with respect to the word forms in the first 200 tokens of a document. It ranges between 1 (high lexical variety) and 0 (low lexical variety)
Family: Morphosyntactic Information	
upos_dist_ADJ	Distribution of adjectives
upos_dist_ADP	Distribution of adpositions
upos_dist_ADV	Distribution of adverbs
upos_dist_AUX	Distribution of auxiliaries
upos_dist_CCONJ	Distribution of coordinating conjunctions
upos_dist_DET	Distribution of determiners
upos_dist_INTJ	Distribution of interjections
upos_dist_NOUN	Distribution of nouns
upos_dist_NUM	Distribution of numerals
upos_dist_PART	Distribution of particles
upos_dist_PRON	Distribution of pronouns
upos_dist_PROPN	Distribution of proper nouns
upos_dist_PUNCT	Distribution of punctuation
upos_dist_SCONJ	Distribution of subordinating conjunctions
upos_dist_SYM	Distribution of symbols

Continued on next page

Table 4: List of features from the "UD Profiling Tool".

Variable	Characteristics
upos_dist_VERB	Distribution of verbs
upos_dist_X	Distribution of other categories
lexical_density	Ratio of content words (nouns, proper nouns, verbs, adjectives, adverbs) over the total number of words in a document
Family: Inflectional Morphology	
verbs_tense_dist_Fut	Distribution of verbs in future tense
verbs_tense_dist_Imp	Distribution of verbs in imperfect tense
verbs_tense_dist_Past	Distribution of verbs in past tense
verbs_tense_dist_Pres	Distribution of verbs in present tense
verbs_mood_dist_Cnd	Distribution of verbs in conditional mood
verbs_mood_dist_Imp	Distribution of verbs in imperative mood
verbs_mood_dist_Ind	Distribution of verbs in indicative mood
verbs_mood_dist_Sub	Distribution of verbs in subjunctive mood
verbs_form_dist_Fin	Distribution of verbs in finite form
verbs_form_dist_Ger	Distribution of verbs in gerund form
verbs_form_dist_Inf	Distribution of verbs in infinitive form
verbs_form_dist_Part	Distribution of verbs in participle form
verbs_num_pers_dist_+3	Distribution of verbs in third person
verbs_num_pers_dist_Plur+1	Distribution of verbs in first person plural
verbs_num_pers_dist_Plur+2	Distribution of verbs in second person plural
verbs_num_pers_dist_Plur+3	Distribution of verbs in third person plural
verbs_num_pers_dist_Sing+1	Distribution of verbs in first person singular
verbs_num_pers_dist_Sing+2	Distribution of verbs in second person singular
verbs_num_pers_dist_Sing+3	Distribution of verbs in third person singular
aux_tense_dist_Fut	Distribution of auxiliaries in future tense
aux_tense_dist_Imp	Distribution of auxiliaries in imperfect tense
aux_tense_dist_Past	Distribution of auxiliaries in past tense
aux_tense_dist_Pres	Distribution of auxiliaries in present tense
aux_mood_dist_Cnd	Distribution of auxiliaries in conditional mood
aux_mood_dist_Imp	Distribution of auxiliaries in imperative mood
aux_mood_dist_Ind	Distribution of auxiliaries in indicative mood
aux_mood_dist_Sub	Distribution of auxiliaries in subjunctive mood
aux_form_dist_Fin	Distribution of auxiliaries in finite form
aux_form_dist_Ger	Distribution of auxiliaries in gerund form
aux_form_dist_Inf	Distribution of auxiliaries in infinitive form
aux_form_dist_Part	Distribution of auxiliaries in participle form
aux_num_pers_dist_Plur+1	Distribution of auxiliaries in first person plural
aux_num_pers_dist_Plur+2	Distribution of auxiliaries in second person plural
aux_num_pers_dist_Plur+3	Distribution of auxiliaries in third person plural
aux_num_pers_dist_Sing+1	Distribution of auxiliaries in first person singular
aux_num_pers_dist_Sing+2	Distribution of auxiliaries in second person singular
aux_num_pers_dist_Sing+3	Distribution of auxiliaries in third person singular
Family: Syntactic Features	
verbal_head_per_sent	Average distribution of verbal heads in the document, out of the total of heads
verbal_root_perc	Average distribution of roots headed by a verb, out of the total of sentence roots

Continued on next page

Table 4: List of features from the "UD Profiling Tool".

Variable	Characteristics
avg_verb_edges	Verbal arity, calculated as the average number of instantiated dependency links (covering both arguments and modifiers) sharing the same verbal head, excluding punctuation and auxiliaries bearing the syntactic role of copula according to the UD scheme
verb_edges_dist_0	Distribution of verbs with arity 0
verb_edges_dist_1	Distribution of verbs with arity 1
verb_edges_dist_2	Distribution of verbs with arity 2
verb_edges_dist_3	Distribution of verbs with arity 3
verb_edges_dist_4	Distribution of verbs with arity 4
verb_edges_dist_5	Distribution of verbs with arity 5
verb_edges_dist_6	Distribution of verbs with arity 6
avg_max_depth	Mean of the maximum tree depths extracted from each sentence of a document. The maximum depth is calculated as the longest path (in terms of occurring dependency links) from the root of the dependency tree to some leaf
avg_token_per_clause	Average clause length, calculated in terms of the average number of tokens per clause, where a clause is defined as the ratio between the number of tokens in a sentence and the number of either verbal or copular head
avg_max_links_len	Mean of the longest dependency links extracted from each sentence of a document
avg_links_len	Average number of words occurring linearly between each syntactic head and its dependent (excluding punctuation dependencies)
max_links_len	The value of the longest dependency link in the document, calculated in number of tokens
avg_prepositional_chain_len	Average value of prepositional 'chains' extracted for all sentences of the document. A prepositional chain is calculated as the number of embedded prepositional complements dependent on a noun
n_prepositional_chains	Total number of prepositional 'chains' extracted for all sentences of the document
prep_dist_1	Distribution of prepositional chains 1-complement long
prep_dist_2	Distribution of prepositional chains 2-complements long
prep_dist_3	Distribution of prepositional chains 3-complements long
prep_dist_4	Distribution of prepositional chains 4-complements long
prep_dist_5	Distribution of prepositional chains 5-complements long
Family: Order of Elements	
obj_pre	Distribution of objects preceding the verb
obj_post	Distribution of objects following the verb
subj_pre	Distribution of subjects preceding the verb
subj_post	Distribution of subjects following the verb
Family: Syntactic Relations	
dep_dist_acl	Distribution of clausal modifiers of nouns
dep_dist_acl:relcl	Distribution of relative clauses
dep_dist_advcl	Distribution of adverbial clauses
dep_dist_advmod	Distribution of adverbial modifiers

Continued on next page

Table 4: List of features from the "UD Profiling Tool".

Variable	Characteristics
dep_dist_amod	Distribution of adjectival modifiers
dep_dist_appos	Distribution of appositions
dep_dist_aux	Distribution of auxiliaries
dep_dist_aux:pass	Distribution of passive auxiliaries
dep_dist_case	Distribution of case markers
dep_dist_cc	Distribution of coordinating conjunctions
dep_dist_ccomp	Distribution of clausal complements
dep_dist_compound	Distribution of compound words
dep_dist_conj	Distribution of conjuncts
dep_dist_cop	Distribution of copulas
dep_dist_csubj	Distribution of clausal subjects
dep_dist_det	Distribution of determiners
dep_dist_det:poss	Distribution of possessive determiners
dep_dist_det:predet	Distribution of predeterminers
dep_dist_discourse	Distribution of discourse elements
dep_dist_dislocated	Distribution of dislocated elements
dep_dist_expl	Distribution of expletives
dep_dist_expl:impers	Distribution of impersonal expletives
dep_dist_expl:pass	Distribution of passive expletives
dep_dist_fixed	Distribution of fixed multiword expressions
dep_dist_flat	Distribution of flat multiword expressions
dep_dist_flat:foreign	Distribution of foreign flat multiword expressions
dep_dist_flat:name	Distribution of names in flat multiword expressions
dep_dist_iobj	Distribution of indirect objects
dep_dist_mark	Distribution of markers
dep_dist_nmod	Distribution of nominal modifiers
dep_dist_nsubj	Distribution of nominal subjects
dep_dist_nsubj:pass	Distribution of passive nominal subjects
dep_dist_nummod	Distribution of numeric modifiers
dep_dist_obj	Distribution of objects
dep_dist_obl	Distribution of obliques
dep_dist_obl:agent	Distribution of agent obliques
dep_dist_orphan	Distribution of orphan elements
dep_dist_parataxis	Distribution of parataxis
dep_dist_punct	Distribution of punctuation
dep_dist_root	Distribution of roots
dep_dist_vocative	Distribution of vocatives
dep_dist_xcomp	Distribution of open clausal complements
Family: Use of Subordination	
principal_proposition_dist	Distribution of principal clauses
subordinate_proposition_dist	Distribution of subordinate clauses
subordinate_post	Distribution of subordinate clauses following the main clause
subordinate_pre	Distribution of subordinate clauses preceding the main clause
avg_subordinate_chain_len	Average length of subordinate chains, where a subordinate 'chain' is calculated as the number of subordinate clauses embedded on a first subordinate clause
subordinate_dist_1	Distribution of subordinate chains 1-clause long

Continued on next page

Table 4: List of features from the "UD Profiling Tool".

Variable	**Characteristics**
subordinate_dist_2	Distribution of subordinate chains 2-clauses long
subordinate_dist_3	Distribution of subordinate chains 3-clauses long
subordinate_dist_4	Distribution of subordinate chains 4-clauses long
subordinate_dist_5	Distribution of subordinate chains 5-clauses long

Table 4: List of features from the "UD Profiling Tool".

Anti-elitism	**Translation**	**People-centrism**	**Translation**
antidemocratic*	undemocratic	abitant*	citizen
casta	caste	cittadin*	citizen
consens*	consensus*	consumator*	consumer
corrot*	corrupt*	contribuent*	taxpayer
disonest*	dishonest*	elettor*	voter
elit*	elite*	gente	people
establishment	establishm*	popol*	people
ingann*	deceit*		
mentir*	lie*		
menzogn*	lie*		
partitocrazia	establishm*		
propagand*	propagand*		
scandal*	scandal*		
tradim*	betray*		
tradir*	betray*		
tradit*	betray*		
vergogn*	shame*		
verità	truth*		

Table 5: Seed words of the dictionary found in Decadri and Boussalis (2020) for anti-elitism and people-centrism.

Variable	**Literature Features**	**Selected Features**
Intercept	−1.00±0.32**	−0.61±0.36
G_index	−0.14±0.11	-
char_per_tok	−0.24±0.13	-
Type_token_ratio	−0.05±0.08	-
word_frequency	0.17±0.10	-
avg_max_depth	−0.05±0.08	-
lexical_density	0.21±0.09*	0.25±0.10**
upos_dist_PROPN	-	0.42±0.09***
Populist_words_ratio	-	0.32±0.08***
verbs_mood_dist_Cnd	-	−0.20±0.08*
verbs_form_dist_Fin	-	−0.43±0.09***
dep_dist_det:predet	-	0.34±0.08***
verbs_num_pers_dist_Sing+2	-	0.24±0.08**
verbal_root_perc	-	−0.19±0.08*
upos_dist_ADJ	-	−0.47±0.10***
subordinate_dist_4	-	0.18±0.08*
verb_edges_dist_1	-	0.19±0.08*
Controls	Yes	Yes

Table 6: Comparative analysis of GLM outputs for literature and automatically selected features with estimates and standard errors. Significance codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Predictor	**Literature Features**	**Selected Features**
Intercept	$-13.90 \pm 1.92^{***}$	$-17.81 \pm 2.77^{***}$
G_index	–0.40±0.31	-
char_per_tok	–0.55±0.38	-
Type_token_ratio	–0.03±0.27	-
word_frequency	0.33±0.27	-
avg_max_depth	0.00±0.21	-
lexical_density	0.13±0.28	–0.37±0.35
upos_dist_PROPN	-	0.32±0.31
Populist_words_ratio	-	0.56±0.32
verbs_mood_dist_Cnd	-	$–0.89 \pm 0.32^{**}$
verbs_form_dist_Fin	-	$–0.93 \pm 0.33^{**}$
dep_dist_det:predet	-	–0.31±0.34
verbs_num_pers_dist_Sing+2	-	0.48±0.28
verbal_root_perc	-	–0.10±0.24
upos_dist_ADJ	-	0.24±0.43
subordinate_dist_4	-	$0.76 \pm 0.31^{*}$
verb_edges_dist_1	-	$0.79 \pm 0.29^{**}$
Controls	Yes	Yes
Random Effects	Variance ± Std.Dev.	
author (88)	342.53±18.51	535.17±23.13
transcriber (11)	0.42±0.65	1.00±1.00

Table 7: Comparative analysis of GLMER outputs for literature and automatically selected features with estimates, standard errors, and random effects. Significance codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Model	npar	AIC	BIC	logLik	Deviance	Chisq	Pr(>Chisq)
model_literature_features	19	451.97	542.06	-206.98	413.97		
model_selected_features	24	416.78	530.58	-184.39	368.78	45.19	1.33e-08 ***

Table 8: Chi-square comparison of the mixed-effects model with predictors derived from the literature and model with the automatically selected features.

Topic by Decade	Key Terms (Italian)	Translation
1990-1999_Topic_0	presidente, governo, paese	president, government, country
1990-1999_Topic_1	governo, paese, presidente	government, country, president
1990-1999_Topic_2	governo, paese, sinistra	government, country, left
2000-2009_Topic_0	governo, paese, presidente	government, country, president
2000-2009_Topic_1	sinistra, liberta, partito	left, freedom, party
2000-2009_Topic_2	anni, parte, governo	years, part, government
2000-2009_Topic_3	governo, partito, paese	government, party, country
2000-2009_Topic_4	legge, referendum, governo	law, referendum, government
2000-2009_Topic_5	politica, lavoro, persone	politics, work, people
2000-2009_Topic_6	sinistra, punto, potere	left, point, power
2000-2009_Topic_7	citta, parte, casa	city, part, house
2010-2019_Topic_0	paese, anni, legge	country, years, law
2010-2019_Topic_1	governo, paese, anni	government, country, years
2010-2019_Topic_2	legge, lavoro, anni	law, work, years
2010-2019_Topic_3	presidente, governo, paese	president, government, country
2010-2019_Topic_4	citta, anni, cittadini	city, years, citizens
2010-2019_Topic_5	lavoro, paese, anni	work, country, years
2010-2019_Topic_6	anni, grazie, paese	years, thanks, country
2020-2023_Topic_0	governo, presidente, ministro	government, president, minister
2020-2023_Topic_1	governo, presidente, ministro	government, president, minister
2020-2023_Topic_2	paese, anni, futuro	country, years, future
2020-2023_Topic_3	regione, persone, anni	region, people, years
2020-2023_Topic_4	ministro, giustizia, signor	minister, justice, mister
2020-2023_Topic_5	lavoro, paese, governo	work, country, government
2020-2023_Topic_6	governo, paese, presidente	government, country, president
2020-2023_Topic_7	legge, parte, anni	law, part, years
2020-2023_Topic_8	presidente, governo, anni	president, government, years

Table 9: LDA topics by decade and key terms.