Nations and Nationalism: A Text Analysis

Vaibhav Jha

2025-12-10

Table of contents

Abstract

This paper analyzes how political discourse varies across nation-states using English-language election manifestos from the Manifesto Project Database. Using collocation networks, it compares how the United Kingdom and France invoke key legal and political concepts, showing systematic differences that align with their common law and civil law traditions. A k-means clustering and principal components analysis of parts-of-speech frequencies reveals cross-national patterns in linguistic style that may shape how political messages resonate with different electorates. Finally, time series of explicitly nationalistic terms suggest that nationalist rhetoric in countries such as the United Kingdom, United States, Turkey, and France has remained relatively stable over the past several decades, complicating popular narratives of sharply rising nationalism. Together, these results illustrate how quantitative text analysis

can uncover both systematic institutional differences and surprising continuities in nations' political self-presentation over time.

1 Introduction

The nation-state is a modern concept. While nation-states have existed for generations, humans organized in many differing ways since the start of the human race. From tribes to empires, human society gave birth to the nation-state following the Early Modern Period (1450-1750). The decline of empires led to the concept of regional sovereignty, where civilizations controlled political, social, and economic life from within their territory. Politics grew to become what we know it to be today, involving grassroots movements and the election of government officials to sustain the nation-state.

Nation-states follow different legal systems, including common, civil, and theocratic law (Barkan 2023). A nation-state's chosen legal system is a reflection of its allies, enemies, and political agenda. While the Occident is often generalized as "the free world," its nation-states largely differ in the intricacies of their legal systems. A comprehensive analysis has the potential to uncover which nations are truly similar in their political discourse and which are significantly different.

This paper explores how political discourse compares across nation-states in the world today. I begin with an analysis of the United Kingdom and France, two prominent Western regimes with differing legal systems. How does word choice in political manifestos compare between a common law legal system and a civil law legal system? Next, I move to a parts-of-speech analysis of 25+ nation-states. Which countries share linguistic similarities in their political discourse? Finally, I study the emotional aspect of nations and nationalism. How has nationalistic sentiment evolved over time? I adopt political manifestos for several countries as a gauge for nationalistic sentiment among a populace.

2 Data

The data for this project is sourced from the Manifesto Project. The Manifesto Project Database was compiled as part of the Manifesto Research on Political Representation Project. It is maintained by the WZB Berlin Social Science Center, viewed as one of the premier comparative data sets in political science.

The authors designed the Manifesto Dataset for political content analyses, later producing a corresponding Manifesto Corpus and the manifestoR package. This allows for accessible text analysis using functionalities provided by the authors to easily filter through and study political manifestos of interest. manifestoR is a free package that allows researchers to access the complete database through an API key.

The corpus consists of 3,000+ documents, each corresponding to a political manifesto from one particular country during one particular election cycle. While the corpus is designed to account for election manifestos across 1,400 political parties from 1945 through today, a subset are available in English. Documents that were either originally written in English or whose

English translation are available were used in this study. This resulted in the use of 1,956 documents (see Table 1).

Table 1: Summary of English corpus

Measure	Value
Number of documents	1,956
Total tokens	2,990,377
Total types (unique words)	242,571
Mean tokens per document	29,904

Each document consists of relevant metadata, including but not limited to a manifesto id, name of party, original written language, and date written. Each date refers to the month/year of the corresponding election cycle.

3 Methods

I began by preparing the data for analysis. This involved filtering for manifestos where the English language was the original written language or an English translation was available. Next, I tokenized the data to create a document-feature matrix. This would allow for collocation analyses and other parts-of-speech studied (described later). Some pre-processing was involved to remove stopwords and other filler words that were not relevant to this study.

The study began with a collocation analysis. To study how legal systems stratify political discourse, I focused on word choice surrounding related topics. I elected to focus on the United Kingdom (U.K.) and France, two prominent nation-states with rich data spanning decades. Prior to studying the frequency of tokens in each nation, I performed a qualitative analysis to brainstorm words of interest to the research question. I noted words that were reflective of legal system discourse, shown in the Appendix, Table A3. Terms like justice, precedent, and trial were included. This step was important so as to not bias my analysis with findings that loosely answered the research question.

I proceeded by scoping the tokens of highest frequency, editing my list of words in consideration accordingly. Next, I found collocations for relevant words in each of the U.K. and France subsets of political manifestos. I identified the top collocates of various words using the MI statistic (3a–MI(0), L10–R10, C5–NC1; PMI >= 6). The PMI (pointwise mutual information) statistic was cut-off at 6 for optimal visualization purposes. I produced collocation plot networks, displaying the collocations for the same nodes between the two nation-states. This allowed for a clear comparison of collocations between differing legal systems.

The second part of my analysis involved a k-means clustering of nation-states by parts-of-speech frequency. My motivation was to discover patterns of parts-of-speech usage between nation-states. This began with a parts-of-speech parsing of the corpus. I used parallel processing techniques to tag each token in the corpus. The corpus was too large to analyze at once following parts-of-speech tagging (7 GB), so the remainder of this analysis demanded a

division of the corpus into four subcorpora of similar length (Brezina 2018). I landed on the use of three centers for each of the k-means clustering algorithms. After evaluating various k values, three produced the most distinct clusters. Following retrieval of frequency summary statistics, I conducted a principal components analysis of the different parts-of-speech statistics. I obtained the top two dimensions and plotted the trends of nation-states alongside the trends of parts-of-speech.

I concluded my analysis with a time series study of nationalistic terminologies. Using the tokens outlined in Table A4, I plotted their frequency distributions over the duration of political manifestos available for the particular nation-state. This side-by-side analysis allowed for nation-state-specific and comprehensive analysis of nationalistic sentiment changes over time. I focused on the nation-states of the U.K., U.S., Turkey, and France due to the richness of data available over a relatively long period of time.

4 Results

The U.K. and France composed a significant part of the corpus, as shown in Table A1 and Table A2. The results of the collocation analyses for the word *justice* are displayed below.

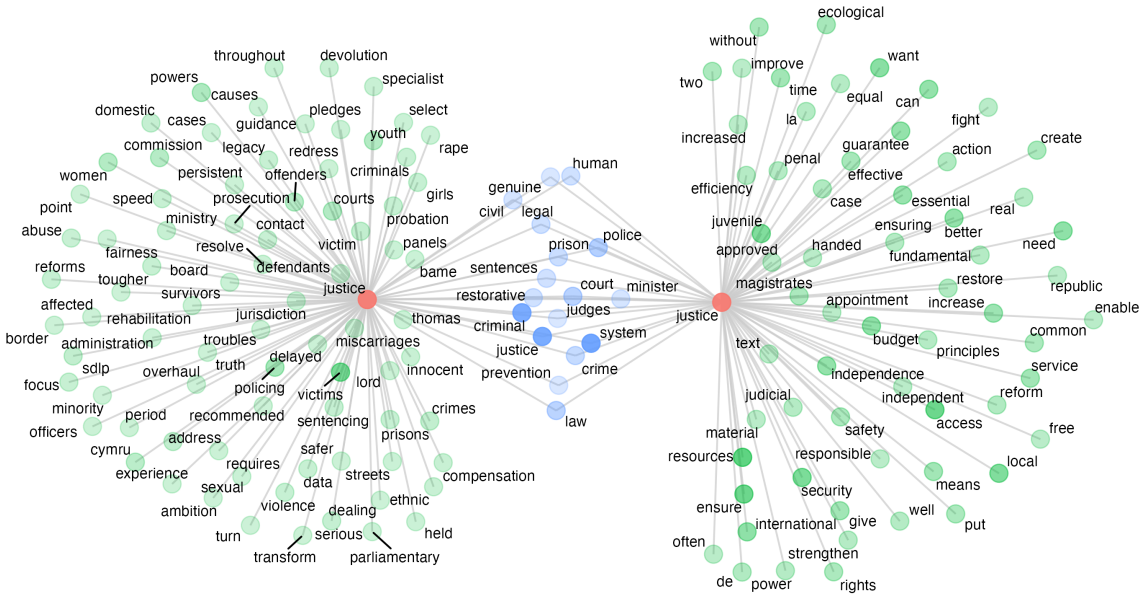

Figure 1: Justice collocation network: U.K. (left) vs. France (right)

The k-means clustering algorithm results were split into four subsets due to the size of the corpus (see Methods). Below is the output of the principal components analysis for the first subset of the corpus by country code and by part-of-speech, split up for visualization purposes.

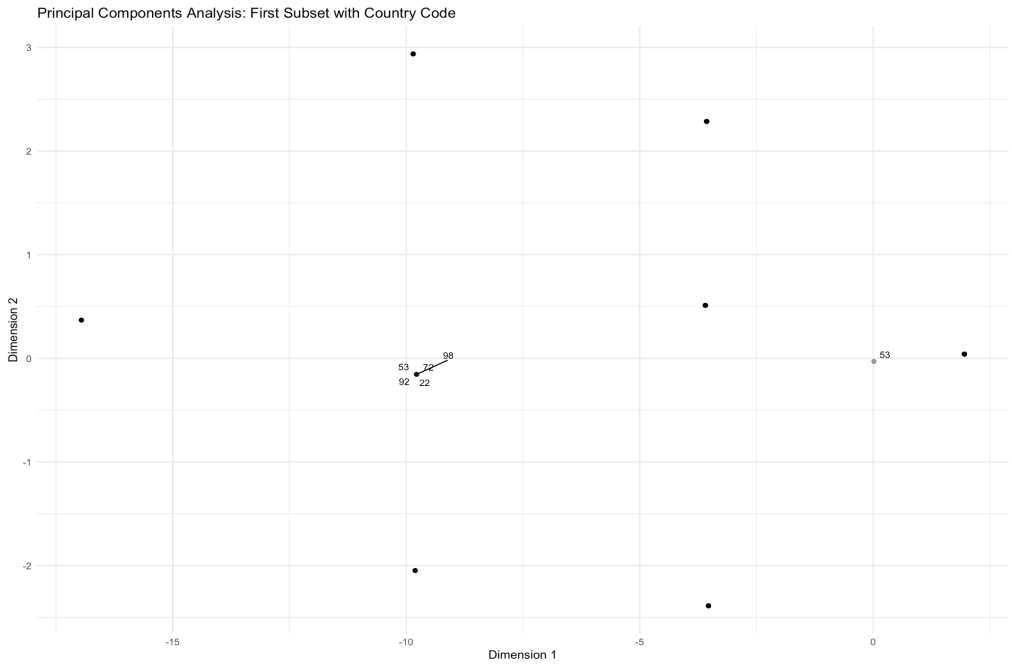

Figure 2: PCA of parts-of-speech by country code (Part 1).

Table 2: Country Code Classification

Code	Country
22	Netherlands
53	Ireland
72	Israel
92	Poland
98	Ukraine

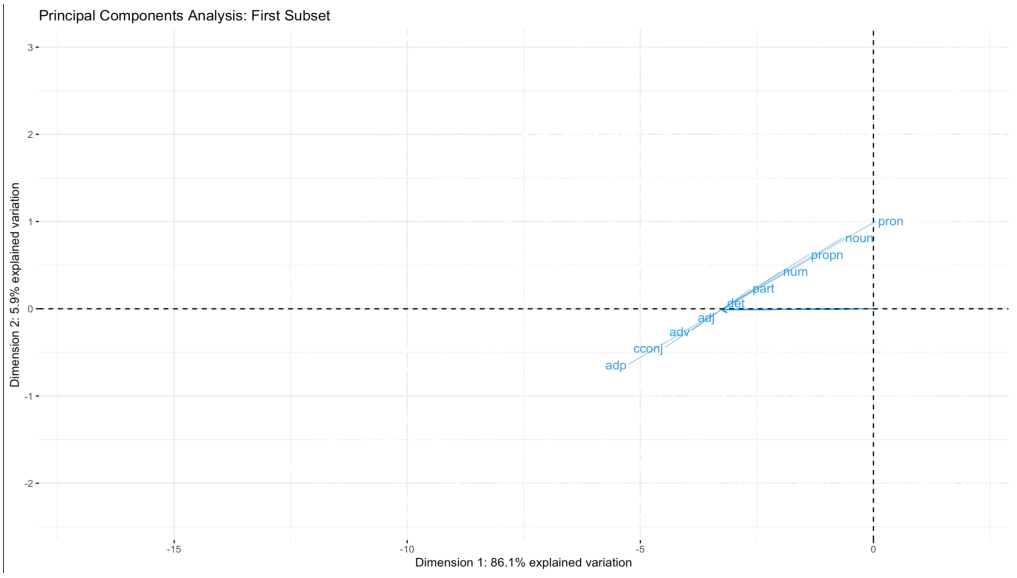

Figure 3: PCA of parts-of-speech (Part 1).

The time series analysis produced frequency trend plots, spanning the duration of political manifesto dates. Below are plots for each of the U.K. and the U.S.

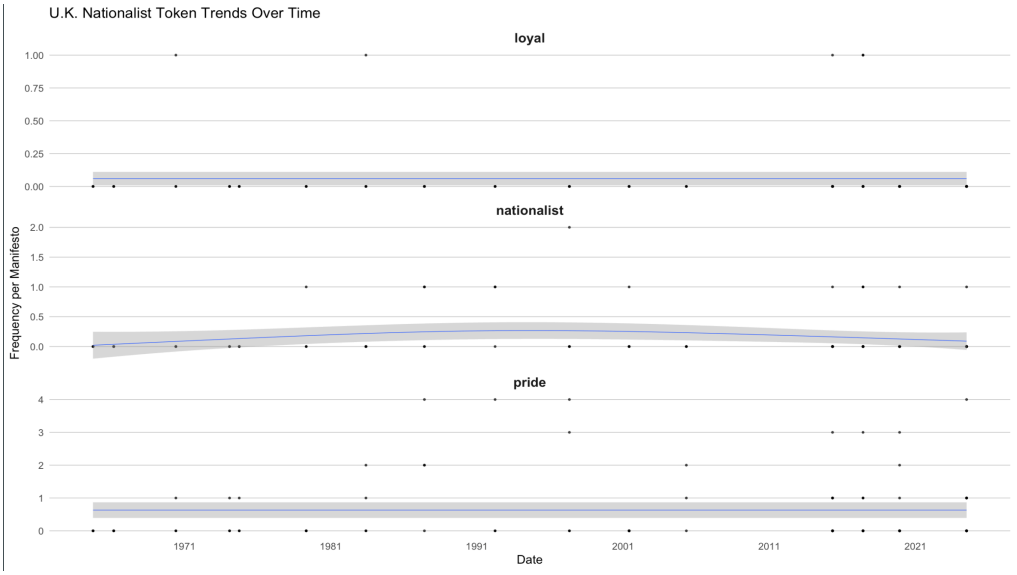

Figure 4: Time series for U.K. manifestos.

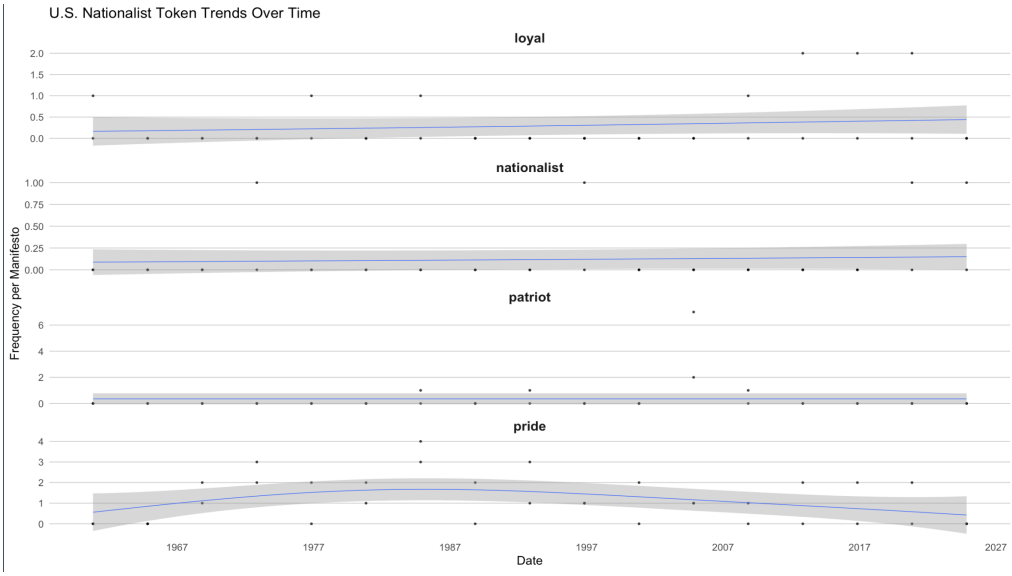

Figure 5: Time series for U.S. manifestos.

5 Discussion

The collocation analysis served to demonstrate how common law countries like the United Kingdom differ in use of legal terminologies in comparison to civil law countries like France. As evident by Figure 1, the common and civil law nation-states share the use of phrases like *criminal*, *judges*, and *prison*, following the theme of justice. The U.K. distinctively referred to justice using terminologies such as *victims*, *offenders*, *policing*, and *miscarriages*, while France uniquely used phrases such as *principles*, *rights*, *republic*, and *magistrates*. When evaluating these trends in the context of the countries' respective legal systems, several thematic patterns arise. Common law nation-states like the United Kingdom use such phrases to emphasize concepts of case handling and precedence. Meanwhile, civil law nation-states like France distinctively use words that appeal to the concepts of case law and systemic guarantees. The collocation plot networks produced for the topics of individuality and equality corroborate these findings (Figure A1, Figure A2). Common law nation-states like the U.K. emphasize precedence, while civil law nation-states value written code to uphold law and order.

The following clustering analysis attempted to zoom out of legal system differences to assess nation-states' use of parts-of-speech in political discourse. The findings were limited for much of the corpus, but the first subset of the corpus did point to some potential findings. The principal components analysis was performed on a subset of countries as listed in Table 2. When paired with the parts-of-speech PCA plot (Figure 3), some nation-states follow distinct patterns in their political discourse. Ukraine was found to use more pronouns and proper nouns in its manifestos, while Poland used more adpositions, coordinating conjunctions, and adverbs. These insights are not valuable when evaluated alone. However, they could be valuable when studying certain nation-states and the efficacy of their political discourse. For example, if a

message resonates more strongly with certain countries compared to others, such an analysis could point to ways for the latter to improve their political messaging.

Finally, the time series analysis provided significant insight as to how nationalistic fervency has evolved over time. While media sources and political leaders often purport feelings of increased national identity in recent time, this analysis showed that there has not been much change in use of nationalistic terminologies over the last 50 years. In evaluating nation-states like the U.K., U.S., and France, the use of nationalistic terminologies such as *loyal*, *patriot*, and *pride* has largely remained consistent across each of the countries' histories of political manifestos. Even Turkey, a nation-state headlining news outlets for conflicts regarding nationalism and Kurdish statelessness, follows a uniform use of nationalistic terminologies over time (see Figure A5). While certain groups may spread an idea of intense nationalism over the last few years, this analysis shows evidence against the concept of increased nationalistic sentiment in contemporary society.

There are numerous limitations to this study, starting with the varying availability of manifestos across nation-states. The differences in documents between nation-states evidently emphasized some political discourses more than others, biasing my analysis. Further, the lack of full English translations for all nation-states in the database limits the internal validity of my study. I aimed to make a global analysis with nation-states spanning all regions of the world, but instead had to focus on the Western hemisphere due to language constraints. Lastly, my lack of context for many nation-states limited my ability to extract meaningful insight from my analyses. I recognize that the qualitative interpretation of my work relies on sound knowledge of historical and current events for each of the nation-states involved (Baker, Brezina, and McEnery 2017).

6 Acknowledgments

I used large language models for numerous instances of debugging R code during my analysis. The size of my corpus lended towards major problems for parsing the text for parts-of-speech. Perplexity helped me solve most of these problems.

7 Works Cited

Baker, Helen, Vaclav Brezina, and Tony McEnery. 2017. "Ireland in British Parliamentary Debates 1803–2005: Plotting Changes in Discourse in a Large Volume of Time-Series Corpus Data." In *Advances in Historical Sociolinguistics*, edited by Tanja Säily, Arja Nurmi, Minna Palander-Collin, and Anita Auer, 83–110. Amsterdam: John Benjamins.

Barkan, Steven E. 2023. *Law and Society: An Introduction*. New York: Routledge.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

A Appendix

Table A1: Summary of U.K. corpus

Measure	Value
Number of documents	84
Total tokens	1,784,358
Total types (unique words)	255,523
Mean tokens per document	21242.36

Table A2: Summary of France corpus

Measure	Value
Number of documents	75
Total tokens	1,152,095
Total types (unique words)	198,993
Mean tokens per document	15361.27

Table A3: Words Involved in the Collocation Analysis

Word
justice
individual
collective
legislation
trial
precedent
liberty

Table A4: Words Involved in the Time Series Analysis

Word
loyal
nationalist
nation
pride

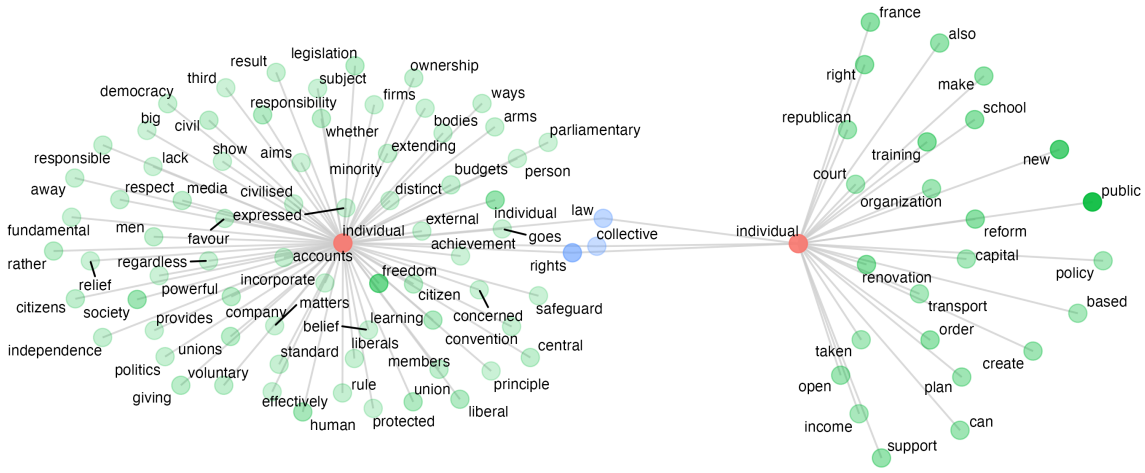

Figure A1: Individual collocation network: U.K. (left) vs. France (right)

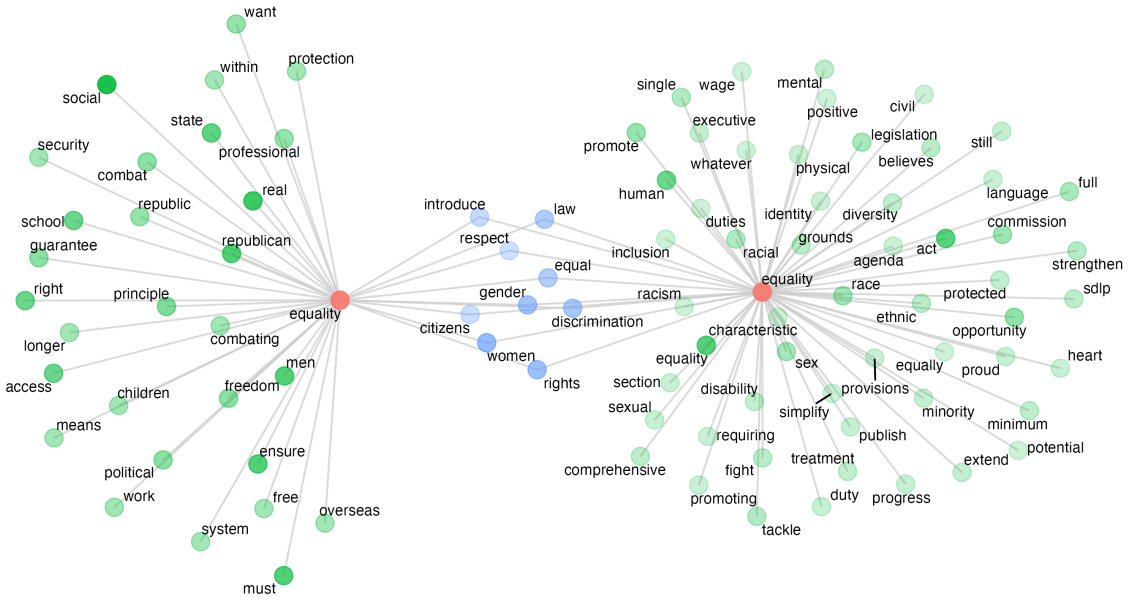

Figure A2: Equality collocation network: France (left) vs. U.K. (right)

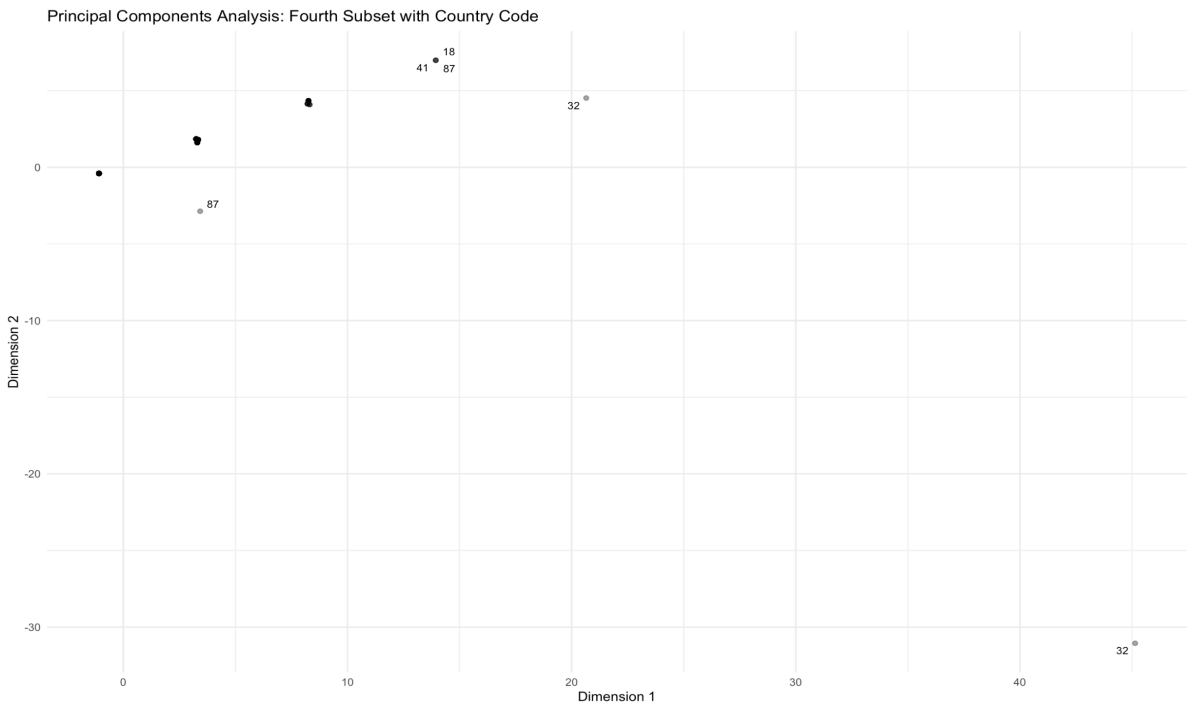

Figure A3: PCA of Parts-of-Speech, by Country Code. Part 4.

Table A5: Country Code Classification

Code	Country
32	Italy
41	Germany
87	Latvia

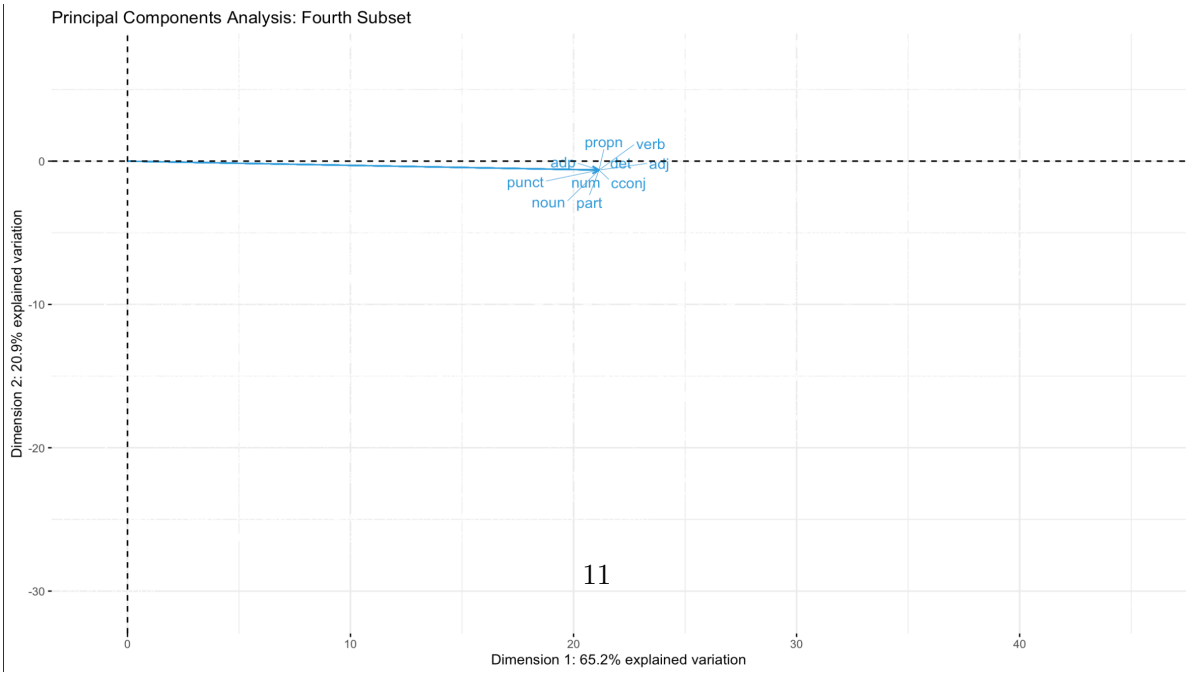

Figure A4: PCA of Parts-of-Speech. Part 4.

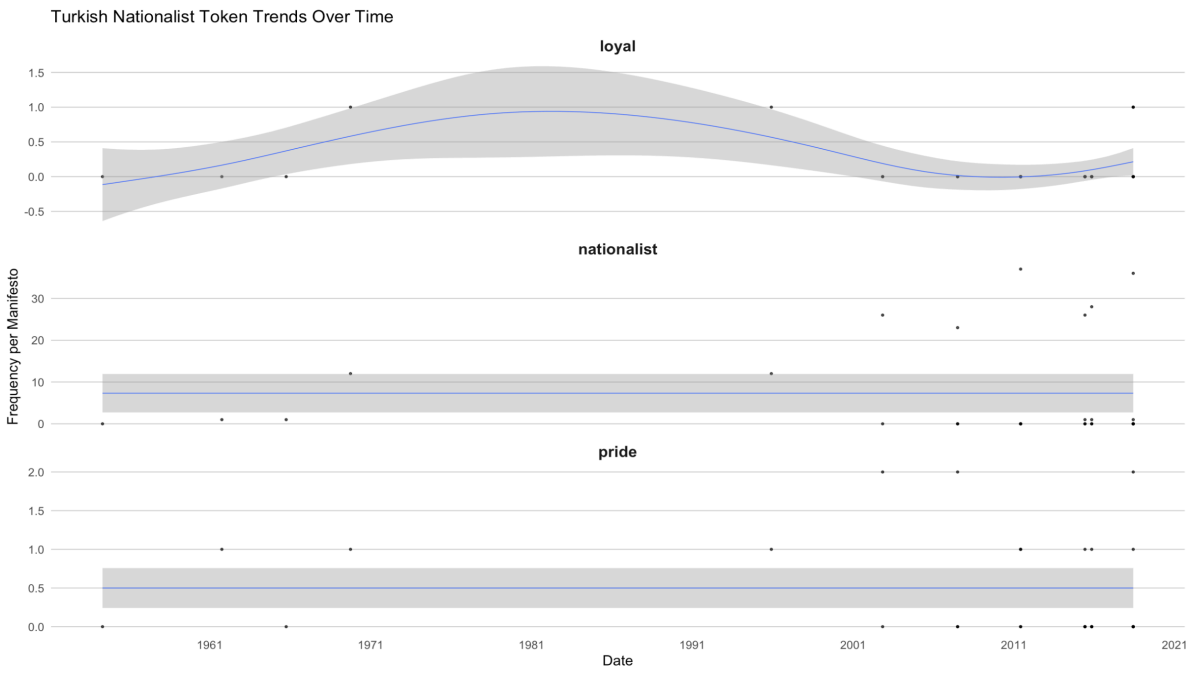

Figure A5: Time series for Turkey manifestos.

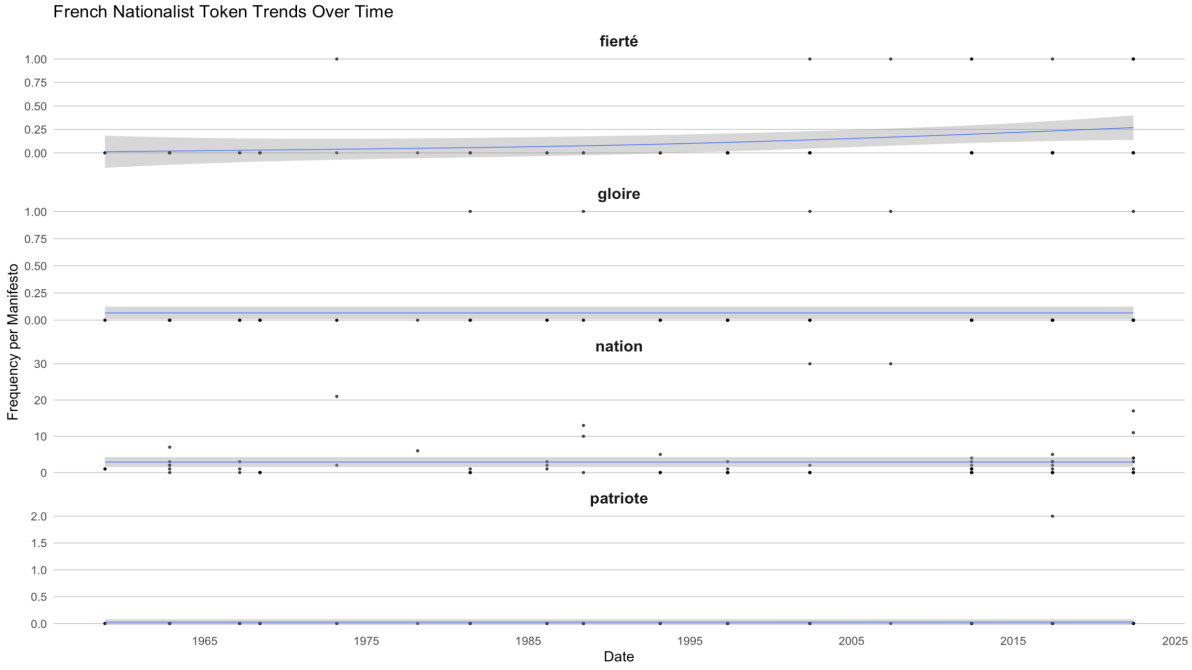

Figure A6: Time series for France manifestos.