BACHELOR'S THESIS / ENGINEERING 2025

Between Words And Power: A Gender-Aware Data Analysis of Spanish Parliament Transcripts

Paula Mateos Marín

BETWEEN WORDS AND POWER: A GENDER-AWARE DATA ANALYSIS OF SPANISH PARLIAMENT TRANSCRIPTS

TREBALL FI DE GRAU DE

Paula Mateos Marín

Directors: Carlos Castillo & Jesús Cerquides

Grau en Enginyeria Matemàtica en Ciències de Dades

Curs 2024-2025

Universitat Pompeu Fabra Barcelona

Escola d'Enginyeria

Als que sempre hi sou

Agraïments

Gràcies als meus pares i els meus avis, pel vostre amor incondicional, per animar-me sempre i per fer-me creure, fins i tot quan jo mateixa no m'ho creia, que podia aconseguir tot allò que em proposés. Sense el vostre suport res d'això hauria estat possible.

A les meves amigues, per escoltar-me quan més ho necessitava, per ser-hi sempre amb paraules d'ànim, i per fer-me riure quan tot es feia costa amunt.

Als meus companys i companyes de la universitat, molts dels quals avui tinc la sort de poder considerar amics. Gràcies per tots els moments compartits, per l'ajuda, les complicitats, les confidències i el suport mutu. Ha estat un privilegi compartir aquest camí amb vosaltres.

Al Bruno, pel suport i pels somriures.

També vull expressar un agraïment especial a totes les persones que m'han acompanyat al llarg de la meva vida acadèmica. A tots els mestres que m'han ensenyat i ajudat a créixer com a persona, des del Sadako, a la UPF passant pel Virolai. En especial als meus tutors de TFG, Jesús i Carlos, gràcies per la confiança, les idees, l'esforç i per creure en aquest projecte des del primer moment. També vull agrair al IIIA-CSIC, concretament al departament de Learning Systems, per l'oportunitat de fer-hi pràctiques, un espai on va néixer la idea d'aquest treball i on vaig poder començar a desenvolupar-la amb el seu suport i col·laboració, però sobretot un espai on he après moltíssim.

Finalment, aquest treball va dedicat a totes les dones que no han tingut veu, però han lluitat perquè avui jo pugui estudiar, escriure aquest treball i decidir lliurement el meu futur. A les que han defensat els drets de totes als carrers, a les institucions, i en tots aquells espais on sovint hem estat silenciades.

Abstract

Sexism in politics is not always strident or visible. Sometimes, sexism manifests in who gets to speak, whose voice is taken seriously, and what words are used to shape power. This study explores how gender inequality has evolved in the Spanish Parliament over the past 20 years. It does so not by simply counting women's interventions, but by listening closely to how they are heard. Through a large collection of parliamentary transcripts, the analysis focuses on participation, language, and silences. It reveals the subtle ways inequality survives even when representation grows. It asks whether symbolic power truly shifts as numbers do, if presence means true power. After all, parliament is meant to reflect society itself. What happens inside its walls does not stay there, it shapes the way we understand voice, authority, and equality both in politics and in everyday life.

Resum

El sexisme en política no sempre és evident. A vegades, es manifesta en qui pot parlar, en quines veus són escoltades amb seriositat i en quines paraules s'utilitzen per construir el poder. Aquest estudi explora com ha evolucionat la desigualtat de gènere al Parlament espanyol durant els últims 20 anys. No ho fa simplement comptant les intervencions de les dones, sinó escoltant de prop com són rebudes. A través d'una àmplia col·lecció de transcripcions parlamentàries, l'anàlisi se centra en la participació, el llenguatge i els silencis. Mostra les formes subtils amb què la desigualtat persisteix, fins i tot quan augmenta la representació. Pregunta si el poder simbòlic canvia realment a mesura que ho fan les xifres, si la presència implica poder real. Al cap i a la fi, el parlament hauria de reflectir la societat mateixa. El que passa dins les seves parets s'hi queda dins, dona forma a com entenem la veu, l'autoritat i la igualtat tant en la política com en la vida quotidiana.

Resumen

El sexismo en política no siempre se muestra de forma evidente. A veces, se manifiesta en quién puede hablar, en qué voces son tomadas en serio y en qué palabras se utilizan para construir el poder. Este estudio explora cómo ha evolucionado la desigualdad de género en el Parlamento español durante los últimos 20 años. No lo hace simplemente contando intervenciones de mujeres, sino prestando atención a cómo son escuchadas. A partir de una amplia colección de transcripciones parlamentarias, el análisis se centra en la participación, el lenguaje y los silencios. Revela las formas sutiles en que la desigualdad persiste incluso cuando la representación crece. Se pregunta si el poder simbólico cambia realmente al mismo ritmo que los números, si la presencia implica verdadero poder. Al fin y al cabo, el parlamento está llamado a reflejar a la sociedad misma. Lo que ocurre dentro de sus paredes no se queda ahí, moldea la forma en que entendemos la voz, la autoridad y la igualdad tanto en la política como en la vida cotidiana.

Index

Introduction....1
1.1 Contextualization.... 1
1.2 Objectives....1
1.3 Hypotheses.... 2
State of the art and background....3
2.1 Previous Studies on Parliamentary Sexism....3
2.2 Computational Approaches to Gender Bias Detection.... 3
2.3 Theoretical Framework.... 4
2.3.1 Legal framing: Insults and Verbal Aggression in Parliamentary Settings....4
2.3.2 Gender-Based Political Violence....4
2.3.3 Non-Sexist Language Guidelines in Parliamentary Contexts.... 5
Methodology....6
3.1 Phase 1: Data Collection and Structuring.... 6
3.1.1 Data Sources....6
3.1.2 Crawling and Preprocessing.... 7
3.1.3 Relational Database Design....8
3.1.4 Enrichment and Cleanup.... 9
3.2 Phase 2: Data and Analytical Preparation....10
3.2.1 Query Data Extraction.... 10
3.2.2 Preprocessing Overview.... 10
3.2.3 Topic Modeling....10
3.2.4 Sentiment Analysis.... 11
3.3 Ethical Considerations.... 12
3.3.1 Use of Public Parliamentary Data.... 12
3.3.2 Handling of Sensitive or Offensive Language....12
3.3.3 Avoiding Bias in Computational Analysis.... 13
3.3.4 Transparency and Reproducibility....13
Results....14
4.1 Political and Institutional Context....14
4.1.1 Political context.... 14
4.1.2 Women's Representation in the Spanish Congress....14
Table 1: Women's Representation in the Spanish Congress.... 15
4.1.3 Gender Composition of the Spanish Government....15
Table 2: Gender Composition of the Spanish Government.... 15
4.1.4 Gender Composition of the Mesa del Congreso....15
Table 3: Gender Composition of the Mesa del Congreso.... 16
4.1.5 Parity by parliamentary group.... 16
Table 4: Parity by parliamentary group....16
4.2 Discursive Participation: Who Speaks and How Much.... 17
Table 5: Total Intervention per Gender per year.... 17
Figure 2: Distribution of interventions by Gender per Year.... 17
Figure 3: Gender Gap in symbolic participation (F% - M%).... 18

4.3 Thematic participation: Who Speaks About What....18
Table 6: Topics by Number of Interventions....18
Figure 4: Gender Intervention Gap by Topic and Year (F% - M%)....19
4.4 Interruption Dynamics....20
Table 7: Interruption rate per gender....20
4.5 Sexist Language....21
4.5.1 When Sexism Is Not Subtle....21
4.5.2 Masculine Generics....21
Table 8: La Población Masculine generic mentions....22
Table 9: La Ciudadanía Masculine generic mentions....22
Table 10: Los diputados y las diputadas Masculine generic mentions....23
4.6 Sentiment Analysis....23
Figure 5: Negative Sentiment Towards Gender (per 1000 words)....24
Discussion....25
Conclusions, Limitations, and Future Work....27
6.1 Conclusions....27
6.2 Limitations....28
6.3 Research transparency....29
6.3.1 Positional Statement....29
6.3.2 Use of AI Tools for Writing Assistance....29
6.4 Future Work....29
Bibliography....31

Chapter 1

Introduction

1.1 Contextualization

In recent years, society has taken visible steps toward addressing gender inequality [1]. Public awareness has grown, and conversations around gender roles, power dynamics, and representation have entered both social and political debate. But has this cultural shift translated into real change within political institutions? Have ideas about equality affected how power is distributed, how voices are recognized, and how participation unfolds in spaces like parliament?

The *Congreso de los Diputados*, as Spain's house of legislative power, comes up as a unique institution from which to study how gender and power intersect within public structures. It is where laws are debated, decisions are made, and the society's democratic will takes shape. As public conversations around gender have evolved, it becomes important to ask whether those shifts are also visible inside this chamber. Parliamentary transcripts offer a unique window into the language of power, into who speaks, how often, about what, and how their words are received. Through that lens, it is possible to explore whether broader changes in gender awareness have truly transformed the discourse at the core of Spain's political system.

1.2 Objectives

This thesis sets out to investigate how gender dynamics are constructed and maintained within the institutional discourse of the *Congreso de los Diputados*. It draws on techniques learned throughout the degree. It adopts elements from feminist theory, discourse analysis, and computational tools to examine the presence, treatment, and symbolic power of women in parliamentary speech.

One central objective is to assess whether the growing presence of women in parliament corresponds to deeper changes in how they participate, how their voices are received, and how power is distributed through language. Rather than assuming that progress in representation implies progress in equality, this study approaches the question through structured and critical analysis. It focuses on participation patterns, thematic associations, tone, and institutional responses during three key legislative years spanning two decades: 2003, 2013, and 2023.

By examining these aspects over time, the project aims to determine whether the observed changes reflect a structural shift or merely a superficial adaptation. The goal is not only to detect possible expressions of sexism, but to evaluate the nature of change itself and consider whether we are moving toward a more equal political space or merely reproducing old dynamics under a new image.

Another equally central objective is to build a centralized, structured, and reusable database of parliamentary discourse. This database is designed to organize fragmented institutional records into a coherent format that enables in-depth, flexible analysis. While it serves the purpose of this study, its potential applications go far beyond gender. The creation of this tool aims to support future research and to contribute to public access, transparency, and data-driven understanding of political speech in Spain.

1.3 Hypotheses

Based on the State of the Art (which will be reviewed in Chapter 2), the following hypotheses guide the project. References below indicate related previous work.

1. Sexism is structurally embedded in parliamentary institutions. Women in parliaments face systematic exclusion, fewer speaking opportunities, less respect, and less power in decision-making just for being women [2, 4].

2. Sexism often appears in indirect or “benevolent” forms. Sexism in political speech often appears in the tone of interventions rather than in explicit hostility [9].

3. Women are more likely to be interrupted, ignored, or mistreated in parliamentary discourse [2].

4. Language reflects and reinforces gendered power dynamics. Word choice, tone, and topic distribution may reveal gender inequalities [8].

5. Numeral representation does not guarantee discursive or symbolic representation. More women do not necessarily mean more influence or power [10,11].

6. During recent years, there has been an improvement in gender representation in political institutions. Despite that progress, there is still a long way to go until Spanish institutions reach true gender equality [1, 11].

Chapter 2

State of the art and background

This chapter has the goal of providing a knowledge base of the previous studies on sexism in the parliament context. Not only through analytical context but also through a more computational approach. By reviewing previous studies on political environments, this chapter will situate the project in the context of similar research with the goal of establishing a starting line for the investigation. Concretely, the study of previous work aims to find gaps and hypotheses to pursue in this thesis. Furthermore, some context regarding laws that regulate *the Congreso de los Diputados* discourses will be reviewed, as well as some theory of gender discrimination.

2.1 Previous Studies on Parliamentary Sexism

As the interest and awareness of gender discrimination have grown in recent years, so have researchers investigating the topic. In the context of parliaments, studies on sexism are guided by testimonies from women in office. For example, the Inter-Parliamentary Union's global report on Sexism, Harassment and Violence against Women Parliamentarians [2] offers a detailed annual study of these issues, supported by regional reports such as the one focused on Europe [3]. These two reports are based on confidential interviews and surveys, and they compute the prevalence rates of sexist behaviour in parliaments around the world. They conclude that political institutions are not free of gender based violence, since disrespectful conduct such as interrupting or ignoring, or direct insults, is common.

Collier and Raney [4] explain that even in institutional environments, some unwritten and unspoken rules and practices generate a sexist environment. By studying real cases and academic literature, they conclude that sexism is normalized and even overlooked in politics. Krook [9], in the same context, explores how political violence often takes place in discourses, with subtle expressions, language helps reinforce gender discrimination without even needing to use a hostile tone.

Despite the significant increase in women's participation in politics, UN Women [10] conclude that women are still underrepresented, especially in the highest level of political power and decision-making. The Gender Equality Index: Spain studies Spain's performance on various gender equality contexts, including politics. It concludes that Spain scores high in formal gender measures and representation, but still shows signs of discrimination in discourses [11].

Together, these studies show that sexism in the parliament is still present., Moreover, it details that sexism is not only the number of women in a chamber, but how these women are treated, what positions they occupy, when, how, and about what they can speak, among many others. However, most of these studies rely on qualitative methods, not analysis of discourses.

2.2 Computational Approaches to Gender Bias Detection

In recent years, as computational methods have evolved and gender bias awareness has increased, so has the application of these techniques to the study of sexism and gender bias. There are many examples of studies focused on detecting sexism, for example, SemEval-2023 [5], which describes the creation of a dataset with over 20000 online comments. Then, it creates a machine learning model to detect sexism. Other works have used fine-tuned models like BERT to detect subtle forms of sexism [6].

Deep learning techniques have also been used to detect sexism. Using LSTM and CNN, Devlin, Gupta, and Shukla concluded in 2021 that deep learning models perform well in identifying patterns

of sexist language in social media [7]. On the other hand, AI models have been found to replicate gender bias even when they were trained to detect it. In this context, the 2024 study titled Explainable AI and Gender Bias, concludes that sexism is so deeply embedded in language that even models trained to detect it fall into the trap of reinforcing sexist power dynamics.

There are many more examples of studies detecting sexism in online platforms like Reddit [12] and YouTube [13], but although these papers provide very interesting points on how sexism can be algorithmically detected, they mainly focus on online content. As far as we know, these techniques have rarely been applied to political debate transcripts.

2.3 Theoretical Framework

The following section will review the rules that shape the discourses in *El Congreso de los Diputados*, since that is the source for parliament speeches that this thesis will use. By understanding these laws and definitions, the conclusions of this research will be more contextualized.

2.3.1 Legal framing: Insults and Verbal Aggression in Parliamentary Settings

The speeches in the Spanish parliament are regulated by the *Reglamento del Congreso*, which defines the norms of conduct and decorum all MPs[1] should follow when they address the chamber. Concretely, the transcripts of the discourses can be affected and even altered by Article 104, which gives the President of the Chamber the authority to intervene when a deputy uses offensive or inappropriate language. The key part that will affect this project is that the regulation allows the President to ask for the removal of the specific offensive comment or sentence so that it does not appear in the *Diario de Sesiones* (*"el Presidente requerirá al Diputado u orador para que retire las ofensas proferidas y ordenará que no consten en el Diario de Sesiones"*) [14].

This practice, while this law was created to protect the order and formality the institution demands, the reality is that it affects discourse analysis, since it biases the data. By removing offensive speech from the official record, the institution performs a discursive erasure that may prevent tracking of sexist or hostile language. As a result, conclusions of studies that analyze sexism and other forms of discrimination may be incomplete or even lead to erroneous conclusions.

2.3.2 Gender-Based Political Violence

Gender-based political violence refers to actions or expressions that aim to diminish, silence, or delegitimize individuals in politics based on their gender. The *Guía sobre la violencia política por razón de género*, developed by HYBRIDAS and the *Dirección General contra la Violencia de Género* in 2021, states that this type of violence can be physical, psychological, economic, sexual, or symbolic [15]. Concretely, in a parliamentary context, symbolic and discursive violence can include persistent interruptions, discrediting remarks, mockery, or even exclusion from debate.

While such behaviors are not reflected in written parliamentary rules, they have a real impact on how women participate and are perceived in institutional spaces. It is not only about being able to speak, it also about how seriously they are taken, the power and weight these interventions are given. This guide emphasizes that symbolic forms of political violence challenge women's authority, credibility, and presence in public life [15].

2.3.3 Non-Sexist Language Guidelines in Parliamentary Contexts

[1] MP stands for "Member of Parliament," referring to elected representatives in the *Congreso de los Diputados*.

The Spanish language traditionally employs the masculine grammatical form as the generic or default. This means that collective nouns such as *los diputados* or titles like *el presidente* are often used to refer to groups or roles regardless of the gender they represent. However, this practice can carry subtle yet persistent forms of symbolic exclusion of women. According to the *Recomendaciones para el uso no sexista del lenguaje*, published by the *Congreso de los Diputados* in 2023, using only the masculine form reinforces gender structures of inequality by making women less visible in the language used inside the institution [26].

This does not go against any formal rule, but it does affect how women are represented and perceived in parliament. The guide suggests more inclusive alternatives, like saying *las diputadas y los diputados* or using *la presidencia* instead of *el presidente*. These changes are not only about grammar, but about ensuring equal space in how political roles are referred to.

Chapter 3

Methodology

This chapter describes the process of methodology that has been followed in this research. It is organized into two phases: the first phase involves gathering, cleaning, and structuring parliamentary transcripts and their metadata into a relational database. The second phase cleanses this data for its use in analysis and employs computational methods to investigate what words were spoken in parliament concerning gender and power structures.

3.1 Phase 1: Data Collection and Structuring

3.1.1 Data Sources

The data for this project was gathered from the *Congreso de los Diputados* official website, using web scraping[2] techniques. Two data sources were used: parliamentary publications and metadata of parliament members.

The scrapping of legislative texts was initially conducted during a research internship at the IIIA-CSIC in 2024. A dataset containing all legislative texts published by the *Congreso* and *Senado* between 1977 and 2024 was generated. Among these texts, there were *Diarios de Sesiones*, legislative texts, financial declarations, and administrative reports, among others. This dataset was then uploaded to Zenodo as "Spanish Congress parliamentary records (1977–2024)" [19], and the code scripts used are open source via GitHub[3]. The web scraping was performed using Selenium WebDriver for browser navigation automation and SQLAlchemy for intermediate data structuring and storage. The metadata of MPs was obtained from the *Congreso*'s Deputies webpage [18], which provides a history of all members who have occupied the chamber and their data, all returned in JSONs. This part of the scraping was programmed this year for this thesis. Getting MPs' data involved interaction with the website's HTML structure, browser inspection, and Selenium automation.

Originally, this scrapping was going to take place using the "All terms" filter on the website, but a problem arose: this filtering option only fetched the latest legislative term of every MP. With that in mind, to avoid loss of information for representatives serving multiple terms, more or less 50% of all MPs, the information was scraped term by term. This approach retrieved a complete historical record that specified precisely who occupied office during each term, but at the price of duplicating entries between terms. MPs who were in office for more than one term were listed multiple times on the database and with different IDs. The procedures used to clean this data are explained in section 3.1.3.

As previously explained, MPs' metadata was retrieved in JSON format, where each response consisted of a data object with an array of entries that mapped to each deputy. Each entry includes fields such as full name, party, gender (numerically indexed), constituency, entry date, exit date, legislative term, among many other fields. This structure was then transformed into a relational database using SQLAlchemy.

Similarly, metadata for parliamentary documents was accessed as JSON and included the URL to the text content, which was later scraped from the HTML pages. This two-step process allowed the scraping to contain all the needed information while being fast and scalable.

---

[2] Web scraping refers to the automated extraction of structured data from websites by simulating user interaction and capturing content embedded in HTML.
[3] GitHub repository: https://github.com/IIIA-ML/CrawlerCongreso

3.1.2 Crawling and Preprocessing

The first scraping of documents conducted in 2024 covered all official *Congreso* and *Senado* publications. However, for this thesis, only the *Congreso de los Diputados Diarios de Sesiones* were retained. They are the only documents including full transcriptions of parliamentary speeches and thus the only ones required for discourse analysis. Furthermore, only publications from 2003, 2013, and 2023 were used. The choice of using only these three years was made to build a working database with the resources at hand, and yet be able to analyze the development of sexism for the past two decades

After the initial filtering, some preprocessing was necessary to make documents consistent, accurate, and ready for analysis. The *Diarios de Sesiones* were extracted from HTML files on the *Congreso* website. While exploring the original dataset, the one generated in 2024, some formatting issues were found. In particular, some simple lines were joined without a space between them, resulting in combined words. Moreover, an encoding issue on the character with code \x97, also known as EM DASH, was found. The particularity of this error was that, in the text field of the database, the character code appeared instead of the dash that was present in the transcripts. These issues were resolved in a custom branch of the scraping repository[4], and the corrected version was used as a basis for all subsequent processing.

Each *Diario de Sesiones* contains an introductory section containing agendas, confirmations, and summaries of attendance. These sections were removed by rule-based scripts so that only the debate section, where MPs speak, remained.

A key part of the preprocessing, and perhaps the most important one, was the segmentation of individual interventions. Using regular expressions with Python's re module, each document was parsed to detect speaker changes and isolate interventions as separate chunks of text. This was possible because the transcripts contained a short introduction of the speaker's identity and position (if any) before the start of each intervention. The intervention segmentation was also built within a Jupyter notebook[5] and enhanced to handle inconsistencies across legislative terms. It was especially built to handle variations in speaker labels, interjections, formatting mistakes, and complicated position titles, among others.

Variability of the introductions before every intervention was one of the main challenges in the segmentation. Usually, intervention headers were in the format beginning with "El señor" or "La señora," which allowed for the first gender classification: interventions launched by "El señor" were labeled with a male (M) designation, and those by "La señora" a female (F) designation, and otherwise labeled as "N". After the formality, the structure would normally consist of either the speaker's last names or their institutional position if any. When a speaker had a title (as in "Ministro de la Presidencia, Relaciones con las Cortes y Memoria Democrática"), the surnames of the speaker would then be listed in parentheses afterwards. The introduction ended with the colon that announces the beginning of the speech. The code was built to work both when a position existed and when it didn't.

Each intervention was separated as an independent text unit and merged with metadata regarding the speaker's surname, gender, document, intervention ID, and legislative term. This design facilitated a speaker-level discourse analysis.

[4] GitHub repository: https://github.com/IIIA-ML/CrawlerCongreso/tree/Paula
[5] GitHub repository: https://github.com/PaulaMateos/Sexism-on-spanish-parliament/tree/main

3.1.3 Relational Database Design

In order to perform a systematic, gender-aware analysis of parliament debate, the scraped and cleaned data were transferred into a relational database in MySQL. The purpose of using this tool was to put the database in a format that would allow for a huge dataset, yet a the same time fast in processing complex queries.

Building the final relational database began by combining two SQLite databases: one contained documents and fragmented interventions, and the other contained MPs by legislative term, along with their metadata.

The original deputy database required restructuring to separate each MP's identity from the legislative terms they served. As previously explained in Section 3.1.1, deputies who participated in multiple terms were initially represented by multiple entries, which created redundancy and inconsistency. To resolve this, two distinct tables were introduced: *diputades*, where each MP is assigned a unique identifier, and *termDiputades*, which stores term-specific information such as party affiliation, institutional role, and constituency, all of which may vary from one term to another. As a result, each MP appears once in the diputades table and is associated with one or more entries in termDiputades using a compound identifier that links the individual to each term served. The *termDiputades* table plays a central role in contextualizing interventions, as it enables changes in political affiliation or institutional role to be accurately tracked across terms, reflecting the nature of parliamentary careers.

Once the identity structure had been mapped, interventions were linked to their speakers' ID by matching the speakers' surnames and gender (metadata available for each intervention) with combinations of surname, gender, and term in the *termDiputades* table.

The majority of entries were successfully matched using this method. However, approximately 5.5% of interventions could not be linked due to either the absence of a compatible speaker profile or the presence of multiple possible matches. These mismatches were typically caused by inconsistencies in the transcripts, such as misspelled speaker names, incorrect or missing gender metadata, or the presence of more than one deputy with the same surname and gender during a given term. For example, Catalan and Basque surnames were often transcribed incorrectly, which complicated automated matching. In some cases, gender inconsistencies were resolved by analyzing contextual clues within the intervention itself. When multiple MPs shared the same surname and gender, further disambiguation was carried out using additional details such as first names presented in parentheses (e.g., “(Don Diego)”) or institutional roles associated with the speaker.

Manual additions to the database were required to include MPs who did not appear in the scraped dataset. This was particularly relevant for ministers who were not officially listed as MPs, as well as for members of Term XV who were no longer in office. The official records for Term XV only include the 350 sitting MPs, so missing individuals were manually inserted to ensure complete speaker coverage. Institutional role data was extracted from the intervention metadata and entered into the position and position_alt columns of the termDiputades table. The most frequently held role for each speaker was assigned to the position field, while any additional roles were stored in position_alt. These additional entries typically corresponded to mid-term role changes or minor transcription inconsistencies, all of which were corrected during the data cleaning process.

The five tables: *document*, *interventions*, *diputades*, *term*, and *termDiputades* resulted from the final database. They were all designed and well-interconnected by the use of foreign keys.

3.1.4 Enrichment and Cleanup

The final preprocessing step involved cleaning intervention texts to enable deeper analysis. Three tasks were carried out at this step: interruption classification, determination of the length of interventions, and introduction cleaning.

To begin with, speaker introductions, those that were key to segment interventions, were removed from the beginning of each intervention, leaving on the text field only the speech content and any potential interruptions in the text field.

Once that was done, interruptions were identified as any content enclosed in parentheses. These were then classified into three categories: positive, negative, or neutral. Positive interruptions include words such as "(Aplausos)" and "(Asentimientos)". Negative interruptions are when there is disagreement with words such as "(Protestas)", "(Rumores)", "(Pateos)", or "(Denegaciones)".
However, not all parenthetical elements were considered interruptions. Some were procedural or structural in nature, such as references to legislative documents or administrative notes (e.g., (número de expediente) or (comienza su intervención en euskera)). These were excluded from the classification, as they did not reflect audience reactions or interjections. Additionally, any interruptions beginning with El señor or La señora were classified as negative, since they typically signaled that another MP was verbally interjecting, often to express disagreement or discomfort.

Finally, some interruptions did not convey either positive or negative sentiment, so they were categorized as neutral interruptions. To capture this classification, three new fields were added to the intervention table in the database: *interruptions*, which record the total number of interruptions per intervention, including unclassified ones, *pos_interruptions*, which store the count of positive interruptions, and *neg_interruptions*, which store the count of negative ones.

The length of each intervention was calculated in characters and added to a field to facilitate length analysis. The measurement enabled the following textual analysis and normalization.

The final version of the structured and organized database, with all those fields that have been outlined hereinabove, is freely available in Zenodo [20]. The following figure is the final database EER model.

Figure 1: EER model

3.2 Phase 2: Data and Analytical Preparation

3.2.1 Query Data Extraction

The analytical phase relied on a well-structured relational database built in Phase 1. This database stored not only each intervention but also its full metadata, including speaker's name, gender, party, date, term, and political role. The goal of this structure was to support flexible queries that could feed each stage of analysis with filtered, clean, and contextualized data.

SQL queries were used to extract subsets of interventions according to specific analytical needs. For instance, discursive participation analysis required a simple query by year and gender to count and retrieve intervention texts, while thematic and sentiment analysis demanded full-text retrieval linked to speaker metadata. These queries also enabled the generation of tables for descriptive statistics, including the total number of MPs and ministers per term, disaggregated by gender. The database design allowed interventions to be easily transformed into CSV files, which served as inputs for subsequent Python-based analysis.

This modular extraction process was crucial to ensure data quality. By working directly from a consistent and normalized source, the risk of duplication or mismatched metadata was avoided. The intervention texts used in every analysis were always grounded in this structured dataset, ensuring replicability and accuracy.

3.2.2 Preprocessing Overview

To prepare the intervention texts for computational analysis, a shared preprocessing pipeline was implemented for both topic modeling and sentiment analysis, ensuring consistency across methods. The primary goal of preprocessing was to reduce linguistic noise and enhance the relevance of lexical features for both thematic and tonal analysis.

The original transcripts were first converted to lowercase to normalize the input. Punctuation, numbers, and special characters were removed, as they carry little analytical value in this context. Common Spanish stopwords were then filtered out using spaCy's built-in stopword list, since these frequent function words (e.g., *pero*, *aunque*, *entonces*) typically do not contribute meaningful content for analysis.

Next, the text was tokenized and lemmatized using spaCy's Spanish language model (es_core_news_sm). Lemmatization grouped different inflected forms of the same word, such as trabajando, trabajó, and trabaja, into a single base form, trabajar. This step helped ensure that concept frequency was measured uniformly across documents.

Unlike typical natural language processing pipelines that retain all parts of speech, this pipeline preserved all alphabetic, non-stopword tokens, without restricting to specific grammatical categories. This approach allowed the subsequent analyzes to remain general and flexible, accommodating both topic modeling and sentence-level sentiment detection without prematurely filtering useful information.

3.2.3 Topic Modeling

The purpose of topic modeling was to identify the main thematic areas discussed in parliamentary discourse and to explore how these themes varied across gender and over time. A Latent Dirichlet Allocation (LDA) model was selected for its interpretability and suitability in analyzing large text corpora in social science research.

After preprocessing, the intervention texts were vectorized using CountVectorizer, with thresholds set to exclude extremely rare or overly common terms. Specifically, words had to appear in at least five documents but in no more than 95% of the corpus to be included in the model. The resulting document-term matrix captured the frequency of relevant terms per intervention and served as input for the LDA algorithm.

Several topic numbers were tested, ranging from 4 to 10. A 9-topic model was selected based on coherence metrics and manual inspection of keyword clusters. Fewer topics tended to merge distinct themes, while more topics created fragmented and less interpretable outputs.

Each topic was labeled manually by reviewing its top 15 most representative keywords. For example, a topic containing terms like *ley*, *igualdad*, *violencia*, and *género* was labeled "Justice and Gender Equality." This manual labeling process allowed abstract clusters to be translated into meaningful categories relevant to political discourse.

The final topic assignments were merged with the original metadata to enable comparisons across gender, year, and political group. This allowed the analysis to trace which themes were more commonly addressed by men or women and to examine whether traditional thematic boundaries were being reproduced or redefined.

The aim of this analysis was not to claim that certain topics belong to one gender or another, but to investigate whether access to different areas of discourse reflects deeper structural patterns of inclusion or exclusion.

3.2.4 Sentiment Analysis

Sentiment analysis was used to evaluate the tone of parliamentary interventions and to identify whether evaluative language patterns varied depending on the gender of the speaker or the gender of the person being referenced.

The analysis was conducted at the sentence level to increase granularity. Texts were first segmented using a Spanish-specific tokenizer trained on the Universal Declaration of Human Rights (UDHR), provided by NLTK's PunktSentenceTokenizer. This training step helped ensure accurate sentence boundaries for Spanish-language text.

Prior to analysis, formal greetings and frequent rhetorical expressions such as "gracias señora presidenta" or "señorías" were removed using regular expressions. These phrases were neutral in tone and occurred frequently, so excluding them helped prevent noise in sentiment scoring.

Each sentence was then processed using TextBlob, which assigns a polarity score ranging from -1 to +1. Sentences were labeled as Negative if their score was below -0.1, Positive if above 0.1, and Neutral otherwise. Additionally, word count was recorded to normalize sentiment data across sentences of varying lengths.

To determine the gender target of each sentence, a custom function checked for the presence of gendered nouns such as *diputada*, *ministra*, *presidenta* for women, and *diputado*, *ministro*, *presidente* for men. These nouns helped identify whom the sentence was referring. Meanwhile, sentiment polarity was mostly shaped by adjectives and other evaluative language that indicated how the subject was being described. Although adjectives were not extracted directly, they are among the main lexical features considered by TextBlob when scoring sentiment.

Aggregated results were then computed per year and per referenced gender. The main metric used was the number of negative sentences per 1,000 words, which normalized sentiment across different levels of speech volume.

To assess model performance, 100 sentences labeled as Negative were manually reviewed for each year. Approximately 40% were found to genuinely express negative sentiment, indicating a high false positive rate. However, since this bias was consistent across years, the results were still valid for identifying relative patterns and group differences over time.

It is important to note that this sentiment analysis was exploratory. It was not used to make claims about individual speakers or specific interventions. Instead, it served to uncover broader tone patterns in institutional discourse and to explore how these patterns may be gendered.
The sentiment analysis was not used to conclude individual speakers. Rather, it functioned as an exploratory tool to understand the tone patterns present in parliamentary discourse and their potential gendered implications.

3.3 Ethical Considerations

This section presents the ethical framework followed throughout the project. Even though the data used is public and the analysis is computational, the research deals with topics like gender, power, and inequality, which are socially and politically sensitive. Ethical care has been applied not only in how data is handled but also in how findings are interpreted and presented. Each of the following sections explains the key ethical decisions made during the project.

3.3.1 Use of Public Parliamentary Data

All the data used in this project comes from the official website of the *Congreso de los Diputados*. These records are public and accessible to anyone, as stated in Article 95 of the *Reglamento del Congreso*, which defines the *Diario de Sesiones* as one of the institution's official publications [27].

Furthermore, the resolution approved by the Mesa del Congreso on June 26, 2012, guarantees public access and permanent digital availability of these documents through the website [28]. Moreover, to comply with the principles of transparency established by Law 19/2013, the *Mesa del Congreso* reaffirmed in 2015 that all individuals have the right to access parliamentary documents, regardless of their format or support, as long as they are part of the institution's public administrative activity [29]. For this reason, these materials can be legitimately used for academic, journalistic, or civic purposes.

Even when working with public documents, this project has taken care to approach them with respect and responsibility. When offensive or sexist statements are referenced, they are not included to single out individual actors, but to analyze how certain discursive dynamics operate within the institutional culture of parliament.

3.3.2 Handling of Sensitive or Offensive Language

Some of the language found in the transcripts includes sexist remarks or offensive expressions. These quotes are presented in their original form, but always with critical explanation. They are not included lightly, and they are not meant to reproduce harm, but to examine how certain discourses operate in formal spaces like parliament.

At the same time, the project is limited by the rules of the institution itself. Article 104 of the *Reglamento del Congreso* allows the removal of offensive remarks from the transcripts, which means that many comments may not even appear in the data. The project takes this into account and treats the material with the understanding that it may already have been filtered and biased by institutional laws.

3.3.3 Avoiding Bias in Computational Analysis

Computational tools help process large amounts of data, but they are not neutral. The decisions made in building the models, such as which words to focus on or how to interpret topics, can introduce bias. This project worked carefully to reduce that risk.

For example, in sentiment analysis, the choice to focus on adjective-noun pairs was based on linguistic reasoning: these are the words that usually carry tone or judgment. In topic modeling, different numbers of topics were tested before choosing the version that made the most sense when read manually.

A manual review of sentiment scores was also done to check the results. Around 40% of sentences marked as negative were negative in meaning. While this shows there are limits to what algorithms can detect, it also supports the overall trends, which were analyzed with caution.

3.3.4 Transparency and Reproducibility

All methods used in this project are documented and based on open-source tools. The scripts used for analysis are available, and the data used is public. Every step, from cleaning the text to generating the models, can be followed and repeated.

The goal was not just to find results, but to make the process clear, so others can build on it or apply it to new contexts. Transparency was important for the project, since without it, the findings of this thesis could be questioned. The findings are not meant to be the final word, but a contribution to the conversation on gender and language in politics.

Chapter 4

Results

In this chapter, all the data collected is rigorously analyzed and visualized. Six different analytical processes have been developed to examine the hypotheses presented in Chapter 1, Section 3. These processes examine the data from multiple perspectives to make sure it is fully understood, and conclusions can be drawn from it.

4.1 Political and Institutional Context

The first step is to understand the political and institutional context of the three years chosen for analysis. Concretely, which parties and people were in power each year, what percentage of MPs of the chamber were women, and what percentage of the government ministries/portfolios were held by women. In this section, the objective is to enlighten the hypothesis that stated that descriptive representation does not guarantee discursive or symbolic representation, hypothesis 5. In other words, to see if presence equals power, not only if women are present, but also how they participate.

4.1.1 Political context

Before diving into actual data, some context of who governed the country each year and what topic was the center of attention is essential to better understand the findings of this chapter.

In 2003, Spain was governed by *Partido Popular* (PP) under President José María Aznar, who governed with an absolute majority. The government's support for the U.S. invasion of Iraq faced overwhelming opposition, with polls indicating that 91% of Spaniards opposed the war [21].

During 2013, the PP, with President Mariano Rajoy, held an absolute majority. Spain was dealing with the aftermath of the 2008 financial crisis. That led to the implementation of strict austerity measures. These policies led to public discontent, culminating in mass protests that criticized economic inequality and political corruption [22].

Finally, in 2023, President Pedro Sánchez from *Partido Socialista Obrero Español* (PSOE) started the year governing in coalition with *Unidas Podemos* in a minority. During this year, concretely the 23rd of July, general elections to the parliament were held, resulting in the re-election of Pedro Sánchez as President of the Government on November 16 of that same year, this time in coalition with *SUMAR* and also in minority. The coalition was established to advance progressive policies, focusing on social and climate justice, labor reforms, and the expansion of civil rights [23].

4.1.2 Women’s Representation in the Spanish Congress

Now that the context is set, it is time to see with actual data if the representation of women in the Spanish parliament has changed in the last 20 years. For that, a table has been generated with the number of MPs who held office during the studied terms. It is important to note that this data is per term, not for that concrete year.

While the Spanish Congress has a fixed size of 350 seats, the total number of deputies serving in a term may exceed this due to replacements that mainly happen because MPs step down from their posts. It is important to note that data for the XV Legislature includes MPs who have served up to April 2025, based on official records and public listings.

Year	Term	Total MPs	Women MPs	% Women MPs	Men MPs	% Men MPs
2003	VII	416	132	31.7	284	68.3
2013	X	437	175	40.0	262	60.0
2023	XIV	407	174	42.8	233	57.2
2023	XV	399	182	45.6	217	54.4

Table 1: Women's Representation in the Spanish Congress

As can be observed on the table, an increase of approximately 8 percentual points in women's representation happened between 2003 and 2013, and between 2013 and 2023 5 percentual point improvement. This data suggests that the increase in female representation is likely not coincidental, but rather a more sustained shift that is slowing down as it reaches full parity.

4.1.3 Gender Composition of the Spanish Government

After analyzing the composition of the chamber, it is essential to understand how this change has been reflected in the national government. It is important to see more women represented, but more representation, in order to be sustainable, must be accompanied by more power.

Year	Term	Total Ministers	Women MPs	% Women MPs	Men MPs	% Men MPs
2003	VII	16	2	12.5	14	87.5
2013	X	13	4	30.8	9	69.2
2023	XIV	23	11	47.8	12	52.2
2023	XV	22	12	54.5	10	45.5

Table 2: Gender Composition of the Spanish Government

It is undeniable that, by the data in the table, the representation of women in the government has increased by more or less 40 percentage points, and this increase has been gradual throughout the last 20 years, increasing year by year, not taking steps back. Furthermore, a milestone has been reached, in the XV legislative term, Spain has more female ministers than male ones. The increase in government representation is bigger than the one in the chamber. In 2003, women represented 31.7% of the MPs, yet only 12.5% of ministers. In 2023, women represented 39.6% of MPs and 54.5% of ministers. It is undeniable that the power of women has increased in the last 20 years.

But still, this table may not reflect the truth behind the power distribution. Not all ministers carry the same amount of power. Portfolios like Home Affairs, Defense, Economy, and Finance are usually the most important ones, to achieve true equality, the distribution of these should also be equitable. In that sense, it is important to remark that in 2023, women not only occupied over half of the ministerial positions, but also key portfolios like Defense, Economy, and Equality. All this data has been extracted from the database generated for this project.

4.1.4 Gender Composition of the *Mesa del Congreso*

The board of the parliament is the institution in charge of moderating the debate, the voting sessions, and enforcing the internal regulations of the chamber. Moreover, the president of the Congress is the third-highest authority in Spain. The following table shows the distribution of power in this institution for the years under study.

Year	Term	Women Members	% Women	Men Members	% Men MPs
2003	VII	2	22.2	7	77.8
2013	X	3	33.3	6	66.7
2023	XIV	4	44.4	5	55.6
2023	XV	5	55.6	3	44.4

Table 3: Gender Composition of the Mesa del Congreso

The representation of women in this organization has grown significantly in the last 20 years. But it is also important to note that in 2003 and 2023 (both 2023 terms), the presidents of *La Mesa* were women. However, this raises concerns about whether such appointments are genuine steps toward gender equality or just symbolic gestures. As Krook (2009) argues, increasing the number of women in political roles does not necessarily translate into real power or influence if institutional power structures remain unchanged. In such cases, female presence may be used to project an image of inclusivity without affecting substantial shifts in decision-making authority [26]. For that purpose, the presidency of the board is an ideal position, since it is very visible, but does not have much power in decision-making.

4.1.5 Parity by parliamentary group

There is now no doubt that numeric representation has improved over the last two decades. Women are now 45.6% of Congress, which means there's still room for improvement to achieve full gender parity. Yet, the improvement is significant, in 2023, term XV, several parliamentary groups almost reach parity or even surpass it. For example, the *Grupo Parlamentario Socialista* has 48.23% of its representatives who are women, and *SUMAR* has 51.85%. *Partido Popular* is also quite close, with 45.26% of women MPs. This is quite significant, since these are three of the four most voted-for forces in the Spanish Congress. On the other hand, the remaining of the four main forces, *VOX*, has only 27.27% of its seats held by women. But from this data, it can be read that the majority of Spanish society believes in gender equality and the importance of parity, at least in a political context. These numbers can be found in more detail in Table 4, which contains the parity by parliamentary group. The source of information is the database created for this thesis.

Parliamentary group	Total	Women	Men	% Women	% Men
Partido Popular (PP)	137	62	75	45.26%	54.74%
Partido Socialista Obrero Español (PSOE)	120	58	62	48.33%	51.67%
VOX	33	9	24	27.27%	72.73%
SUMAR	27	14	13	51.85%	48.15%
Esquerra Republicana (ERC)	7	4	3	57.14%	42.86%
Junts per Catalunya	7	2	5	28.57%	71.43%
EH Bildu	6	3	3	50.00%	50.00%
EAJ-PNV	5	1	4	20.00%	80.00%
Grupo Mixto	8	4	4	50.00%	50.00%

Table 4: Parity by parliamentary group

4.2 Discursive Participation: Who Speaks and How Much

The second step in this process to deep dive into the gender and power dynamics of the parliament is to stop counting MPs and see who speaks and how much. MPs are in the chamber to represent the population, and power is given to them by society, but to exercise this power, it is important not only to vote but also to speak one's mind. The goal of this second analysis is to revisit hypotheses 1 and 5, since hypothesis 1 states that sexism is structurally embedded in the parliament institutions, and hypothesis 5 states that descriptive representation does not guarantee discursive representation.

With that in mind, the following is a table counting the number of interventions per year per gender. For this section of the project, interventions by the members of the board of the chamber will not be counted. That decision has been made in order to preserve the true nature of the data, since before any intervention, the president of the board would present the speaker, which means that for every intervention there is another one that lacks content, since it only has the objective of moderating the debate.

Year	Total nº of interventions	Female interventions	% Female interventions	Male speakers	% Male speakers
2003	6329	1200	18.96	5129	81.04
2013	5582	1769	31.69	3813	68.31
2023	2909	1217	41.84	1692	58.16

Table 5: Total Intervention per Gender per year

The findings derived from this table are quite significant. In 2003, women were roughly 30% of the MPs, but only made up 19% of the interventions, which means that while present, most of them did not have a voice. A similar thing, though not that extreme, happened in 2013. On the other hand, in 2023, women represented more or less, since it is a two-term year, 40.5% of the chamber and also 41. 84% of the interventions, indicating that the proportion between deputies' gender and interventions by gender was balanced. It is also noticeable that this balance has significantly increased over the last two decades, as it can be seen in Figure 2.

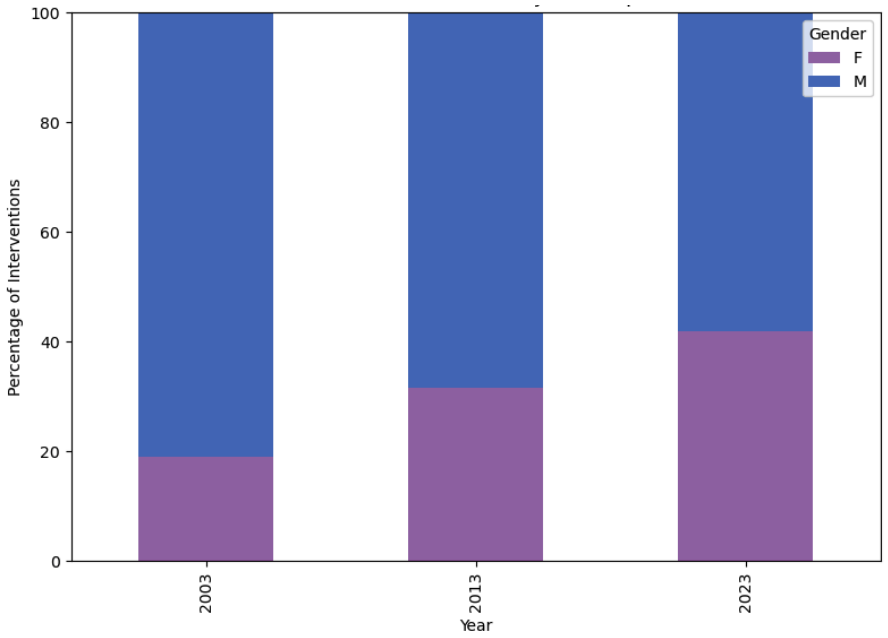

Figure 2: Distribution of interventions by Gender per Year

Having this dataset also allows us to explore how many of the representatives participate in the debate. MPs, by default, will always intervene once to swear their charge; that way, it is possible to classify MPs between active and inactive, where they are active if they have intervened at least twice in the debate. The following heat map represents the findings.

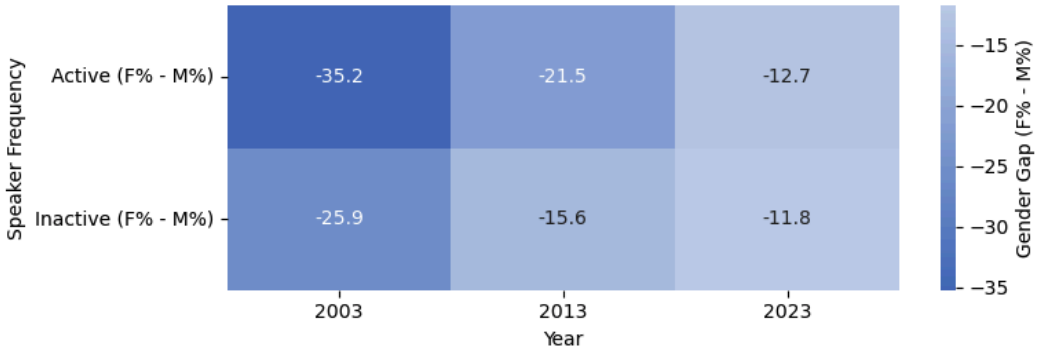

Figure 3: Gender Gap in symbolic participation (F% - M%)

The above figure shows the difference between female active and male active MPs. The figure indicates, from the active MPs, the difference in percentage between male and female. If the distribution among genders of active MPs were fair, the difference in percentage points should be the same for active and inactive deputies. This condition for fairness is only achieved in 2023, which means that in 2003 and 2013, women were not only significantly less present, but they were also even less active. Basically, the already few female representatives were more likely to be inactive than their male counterparts.

4.3 Thematic participation: Who Speaks About What

The third step of this analysis is to go even deeper into the interventions, now not only to count them but to understand their content. Concretely, the goal of this section is to understand if the topics MPs speak about are determined by gender, if there are traditionally considered “women's topics” and “male topics”. A key question here is whether women are relegated to “secondary topics”? This exploration arises from hypothesis number 4, which states that language reflects and reinforces gendered power dynamics.

To try and prove or enlighten the hypothesis, a topic modeling over the interventions was conducted. The most spoken topics in the parliamentary debate, in descending order, were found to be:

Topic Name	Number of interventions
Political Debate	4264
Parliament Procedures	2415
Budget and Public Policy	2324
Justice and Gender Equality	1163
Foreign Affairs and Defence	1158
Autonomous Communities	1095
Employment and Labor	1021
European Union	783
Education and Cultural Policy	597

Table 6: Topics by Number of Interventions

And with that, the interventions were classified per topic. To better capture the result, a heatmap, resembling the mechanics used in section 4.2, was created. Once again, the number of interventions per topic was classified by gender, and then the percentage of interventions of a topic by male speakers was subtracted from the percentage of female speakers. Furthermore, since the output is a heatmap, the darker the blue of a cell, the more male-dominated the topic, the same applies to purple and female-dominated topics.

Figure 4: Gender Intervention Gap by Topic and Year (F% - M%)

This figure is perhaps the most significant visual representation of how gender dynamics are shifting toward equality. It is clear that in 2003, all topics were overwhelmingly dominated by men, which may make sense given that men made up 70% of the chamber. However, some numbers still do not add up. For example, employment and labor (Topic 4) or foreign affairs and defense (Topic 6) show a 70% difference between male and female speakers, which, through a basic proportion calculation, reveals that approximately 85% of the speakers in those categories were men. In fact, for a category to be in balance with the gender distribution in 2003, the difference should be -40 (30-70). With that in mind, not a single topic comes close to the balance point. Nevertheless, the education and cultural policies topic is overrepresented by women, which could suggest it was considered a female topic.

In 2013, at first glance, a slight improvement can be observed, the cells are a lighter shade of blue, which indicates less male dominance. But as with any data, context is required. In 2013, 40% of MPs were women, which means that for a perfect balance in proportion to the congress gender distribution, the difference should now be -20 (40-60). Unfortunately, only two categories come close to that number: employment and labor, and budget and public policy, which is interesting, since those were highly male-dominated topics in 2003. On the other hand, it is in this year that the deepest shade of blue is found, in foreign affairs and defense, a topic where only 12% of speakers are women.

Finally, in 2023, a very notable change can be observed. Not all cells are blue anymore, now there are white ones, and even some purple. This shift is especially worth highlighting when considering that the proportion of women MPs in Congress remained the same as in 2013. As of 2023, 5 out of 9 topics are, in proportion with the distribution of the parliament, female-dominated. Moreover, one of those topics, justice and gender equality, is dominated by women, 60% of speakers are women.

4.4 Interruption Dynamics

Following the sexism analysis, the next step is to investigate hypothesis 3, which states that women are more likely to be interrupted, ignored, or mistreated in parliament discourse [2]. For this concrete section, focus will be placed on the interruptions, since, as explained in the methodology, interruptions in parliamentary discourse transcripts are relatively straightforward to identify and classify. Moreover, for this section, all interventions, including those of chamber members to moderate the debate, will be used.

The main objective of this section is to explore whether women are more interrupted. But concretely, if they are more negatively interrupted, listeners feel more comfortable loudly expressing their discomfort when a woman is speaking. But since men speak more than women as observers in previous sections of this chapter, to make a fair comparison of interruptions, a rate of interruptions per 1k words was created. This way it can be studied in the same number of words, which is more interrupted. The result is expressed in *Table 7*.

Year	Gender	Positive interruptions	Negative interruptions	Positive interruptions per 1k words	Negative interruptions per 1k words
2003	F	525	923	00.9	0.16
	M	1566	1691	00.9	0.09
2013	F	928	517	0.14	0.08
	M	1983	1108	0.12	0.07
2023	F	1115	623	0.22	0.13
	M	1172	700	0.17	0.10

Table 7: Interruption rate per gender

This table was directly generated through a query on the dataset using the interruptions and length fields on the interventions table. At first glance, every year under study, women are more interrupted than men. That is especially clear in 2023 and 2003, but it is also important to note that women are also more positively interrupted than men are.

The values in the table are very small because they are rates. For that reason, it makes sense to do hypothesis testing to truly understand if the difference is statistically significant. To assess that, a one-sided paired t-test was conducted, since with that method, there is no need for individual standard deviations of the populations because what matters is the differences between the pairs (F-M). The null hypothesis that men and women are equally negatively interrupted was rejected with a confidence interval of 90%, but the difference is not statistically significant at the 95% confidence level.

Similar results were found for positive interruptions, there is significant statistical evidence to reject that women are equally positively interrupted in favor of them being more interrupted at a 90% confidence level, but not at a 95%. The p-values obtained for both tests were 0.087 for negative interruptions and 0.125 for positive ones, which means the difference between men and women is more significant in negative interruptions. Overall, the conclusion is that women are more interrupted than men.

4.5 Sexist Language

4.5.1 When Sexism Is Not Subtle

Towards this project, the focus when studying sexist behaviours has been on the most subtle forms of it. But it is reasonable to think that, since so many signs of embedded sexism in the parliamentary institution have been found, there will be some clearer examples, like direct insults towards women. The truth is that very few sexist insults or insults at all are recorded in the transcripts. Due to the already discussed article 104 in the *Reglamento del Congreso*, any offence is removed from the transcripts prior request from the president of the chamber.

Regardless of the rule, it seems like women representatives have lately been speaking up and asking for the insults towards them to remain on the transcripts, so it is engraved in history and no one can deny the existence of this mistreatment. No examples of this groundbreaking behaviour of women were found in 2023 transcripts, but a few were found in prior and subsequent years.

To begin with, and perhaps the most denigrating example, on November 23rd, 2022 [24], Ms. Toscano De Balbín from VOX, dedicated in her speech, the following words to the Minister of Equality, Ms. Montero Gil: *"el único mérito que tiene usted es haber estudiado en profundidad a Pablo Iglesias"*, which translates as "The only merit you have is having studied Pablo Iglesias in depth", stating that her only merit for becoming minister was being the partner of a male politician. A part from the undeniable sexism in this words that tried to diminish the minister's credibility by reducing her to a simple extension of her partner, perhaps the most concerning part is that this words where part of a speech, not an interruption in the heat of the moment, but a rehearsed and calculated attempt to publicly humiliate her, not based on her actions or policies, but on her personal life.

Unfortunately, the examples do not end there, on February 22nd, 2024 [25], when Ms. Martínez Barbero finished one of her interventions, Mr. García Gomis, once again, a VOX representative, said: *"Madre mía, la chavala. Nos lo vamos a pasar bien"* which loosely translates as "Oh my god, the girl. We're going to have fun.". While this comment is not as clearly sexist as the previous one, the message is deeply disrespectful. Referring to a woman MP as "la chavala" reduces her to a girl, framing her not as a peer or authority figure, but as someone young, inexperienced, and out of place. The phrase "madre mía" adds a tone of mockery or exaggeration, as if her intervention were irrational or amusing. But perhaps the most troubling part is "nos lo vamos a pasar bien," which frames her political participation as entertainment. Instead of responding to the content of her speech, the speaker turns her presence into a spectacle, suggesting she will be a source of drama or chaos.

These are only examples, but the list goes on, and these are only the cases that are reflected in the transcripts. One can ask how many of these comments women have had to endure since 1977 that have been erased from history. Article 104 of the *Reglamento del Congreso*, while originally intended to preserve decorum and respect in the chamber, has ended up acting as an accomplice in the perpetuation of sexist behaviours by erasing any trace of them.

4.5.2 Masculine Generics

The Spanish language tends to use the masculine form of nouns as generic, especially in the plural form. This practice is an inherited form of sexism that the language has, every Spanish speaker has learned the language using these generic forms, making it very difficult to implement changes. Nevertheless, the parliament is an institution that represents all Spaniards, and to do so, it must be more formal, and representatives should make an effort to use non-sexist language. It is with this purpose that the Mesa del Congreso released in 2023 a guideline with recommendations for MPs not to use sexist language, as seen in section 2.4.3. This guide was released in December 2023, so it makes no sense to try to find out if the recommendations were followed.

Regardless, what can be done is to analyze the examples the guide uses to find if, even before the recommendations, something was changing among the MPs' use of vocabulary. This can be done by querying the database.

To be more precise, the guideline [16] provides several examples of sexist language in the use of adjectives, nouns, articles, and pronouns, and offers recommendations on how to replace them with non-gendered alternatives. But the guide also recognizes that there are cases where these alternatives do not exist or are too complex.

To analyze the evolution of the use of sexist language, a keyword search has been done and normalized over the total length of all interventions in a year. Since not all years have the same number of documents, this provides a fair comparison without a basis. There are many examples to choose from, but in this parliamentary context, it makes sense to examine the use of the following nouns and their alternatives.

Year	% "Los españoles" mentions	% "Las españolas" mentions	% "La población" mentions
2003	0.0415	0.0010	0.0023
2013	0.0590	0.0014	0.0021
2023	0.1607	0.0050	0.0028

Table 8: La Población Masculine generic mentions

For example, when referring to Spaniards, the guide recommends using *"La población"* (the population, a gender-neutral expression) instead of *"los españoles"* (the spaniards, refering to men). This is one of the simplest alternatives to implement, as it requires no change of phrasing or sentence structure, and the substitute is not longer than the original sexist expression. Table 8 shows that mentions of *"los españoles"* have increased way more than the *"las españolas"*(the spaniards, refering to women) or *"la población"* mentions. This suggests that in the parliament, not many MPs were making an effort to appeal to all Spaniards regardless of their gender.

Year	% "Los ciudadanos" mentions	% "Las ciudadanas" mentions	% "La ciudadanía" mentions
2003	0.1142	0.0018	0.0179
2013	0.0590	0.0014	0.0422
2023	0.0810	0.0032	0.0491

Table 9: La Ciudadanía Masculine generic mentions

The second example is very similar to the first one, again, a very simple substitution, *"los ciudadanos"* (the citizens, refering to men) for *"la ciudadanía"* (a gender-neutral alternative), still the same noun, but with a different conjugation. The results in Table 9 are very similar to the previous table. Again, *"los ciudadanos"* is still the most used form of the ones in the study, but less than in 2003, and *"las ciudadanas"* (the citizens, refering to women) and *"la ciudadanía"* mentions have skyrocketed since then. Still, the difference is big, but at least some improvement seems to be going on, the mentality seems to be shifting.

Year	% “Los diputados” mentions	% “Las diputadas” mentions	% “Las diputadas y los diputados” mentions
2003	0.1142	0.0006	0.0018
2013	0.0590	0.0007	0.0043
2023	0.0430	0.0022	0.0094

Table 10: Los diputados y las diputadas Masculine generic mentions

Finally, the MPs themselves, as seen in previous sections, the parliament is nowadays close to gender parity, but still, the comparison between “*los diputados”* (the male deputies) and “*las diputadas y los diputados”* (male and female MPs) is significant. It is important to note that, since there are many variations for referring to both genders of MPs, all of them have been included in the count represented by “*las diputadas y los diputados”*. This includes phrases like “*los y las diputadas”*, “*las diputadas y diputados”*, and the reverse structures such as “*los diputados y las diputadas”*. Regardless of all the alternatives included in the keyword search, the rates obtained are still far from the problematic phrase. Despite this, it is important to observe that these differences have reduced in the last 10 years, and it seems like the awareness is growing, since the mentions of both genders' alternative is growing.

To enhance the analysis, a search was conducted focusing only on both-gender alternatives that specifically include “*las diputadas”* (the female deputies). This revealed that *“las diputadas”* is mostly mentioned within inclusive expressions, rather than on its own. But what can be expected from an institution called *El Congreso de los Diputados*? It seems hypocritical to ask MPs to use gender inclusive language when the institution itself uses only one gender in its name.

4.6 Sentiment Analysis

The last step of this analysis is sentiment analysis, guided by hypothesis number 4 that states word choice, tone, and topic distribution may reveal gender inequalities. Sometimes a sentence or remark is not offensive because of the content, but because of the intention, the tone, or the specific word choice. Unfortunately, this analysis is very difficult with just transcripts, since they do not reflect the tone or the delivered looks, or the pauses. But what can be done is to explore if the sentiments reflected in the transcripts are different depending on who the target is.

As explained in section 3.2.2, to prepare for this exploration, the data was segmented by sentence, and the adjective and noun were identified, since that is a way of knowing who the receiver of the words is. The following figure shows the evolution of negative sentiment towards women per 1,000 words. This rate of 1,000 words is set to even out the number of sentences that refer to women and men, so the results are fair.

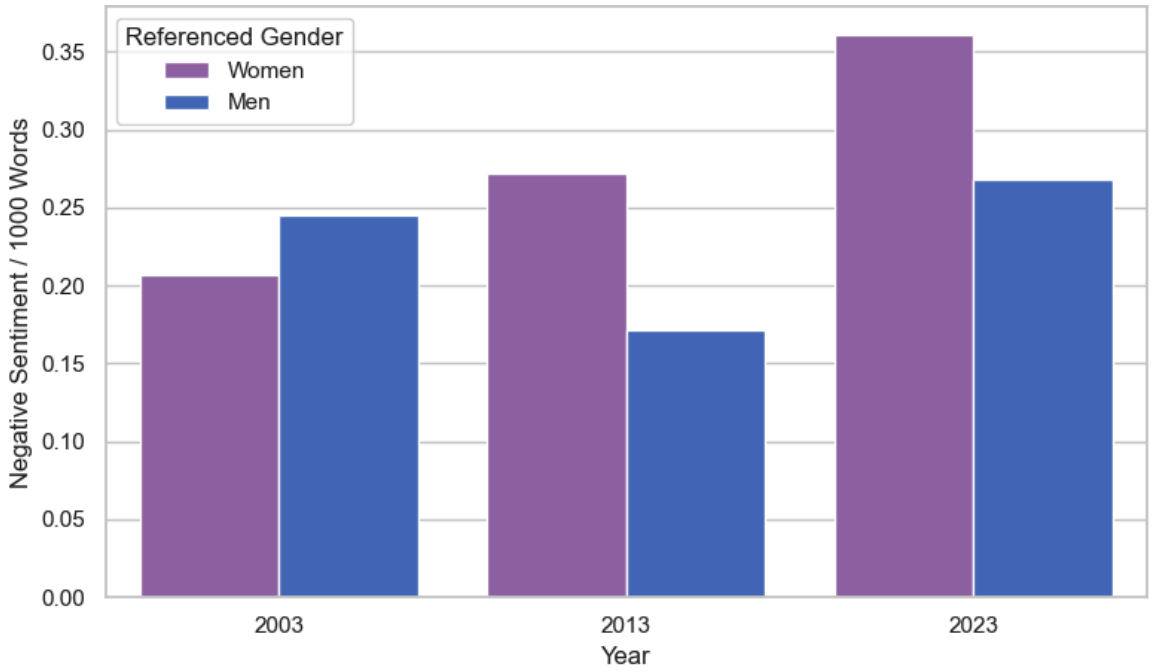

Figure 5: Negative Sentiment Towards Gender (per 1000 words)

As shown in Figure 5, negative sentiment towards women has increased over the last two decades, reaching its highest point in 2023. This contrasts with the patterns observed in previous sections, where indicators of sexism seemed to have improved slightly, suggesting a slow but measurable reduction in gender bias. This apparent contradiction raises some important questions. Could it be that, as women take on more prominent roles within the institution, including leadership positions and important portfolios, they also become more frequent targets of criticism or hostility? If so, this would align with the idea that increased visibility does not always lead to increased legitimacy. On the contrary, it might provoke stronger reactions from those who still believe political authority is a male space.

Of course, the limitations of transcript-based sentiment analysis must be taken into account. As discussed earlier, tone and irony, as well as delivery, are not captured in text, which makes sentiment classification subjective. To ensure the reliability of this step, a sample of 100 sentences labeled as negative was manually reviewed for each of the three years. The outcome of this exploration suggests that only around 40% of these cases could be described as having negative sentiment. This means that, although the model applied provides useful insights, the results must be taken with care.

Chapter 5

Discussion

After presenting the data from six different analytical lenses, this chapter brings the findings together to reflect on the implications of gender dynamics within the *Congreso de los Diputados*. Instead of revisiting individual numbers, the goal is to interpret the results as a whole, linking trends and patterns across the analyzes and exploring what they reveal about the state of gender equality in Spanish parliamentary discourse.

First and foremost, the results show that gender parity in presence is closer than ever, across political groups and institutional structures, and the representation of women has increased over the past two decades. This includes not only the composition of the chamber but also ministerial positions and the board of the Congress. Yet, despite these gains, presence has not always equated to equal influence. The thesis hypothesis that descriptive representation does not guarantee symbolic or discursive representation remains a valid concern.

A particularly relevant insight emerges when participation is considered beyond simple numeric presence. The data reveals a consistent trajectory in which women have progressively claimed more speaking time and thematic space. The evolution from 2003 to 2023 is not only quantitative but qualitative: women are no longer relegated to topics associated with so-called “traditional roles” of women but now lead discussions in areas that were once heavily male-dominated. Topics such as Justice and Gender Equality are no longer marginal but central to parliamentary debate, and they are increasingly shaped by female voices. This shift is a strong signal that gendered barriers in access to discourse are weakening. However, the persistence of some gendered topic associations, particularly in areas like Foreign Affairs and Defense, while still slightly male-leaning, no longer reflects rigid gender boundaries. Overall, the distribution of thematic participation is now much more balanced, indicating that traditional gendered topic dynamics have largely faded.

These structural improvements coexist with forms of subtle and explicit resistance to gender parity. The analysis of interruptions showed that women face more reactions to their speech than men, both positive and negative, which points to a discursive environment in which their authority is more frequently challenged. Similarly, sentiment analysis revealed that negative tone directed toward women remains disproportionately high, even at a time when their institutional presence is at a historic peak. This contradiction supports the hypothesis that visibility alone does not secure power. Greater visibility may even provoke stronger pushback from those who feel threatened by changing norms.

Such pushback is not merely statistical or symbolic but sometimes verbal and direct. Though rare in the transcripts due to article 104 of the *Reglamento del Congreso* [14], explicitly sexist comments continue to appear. While some are removed due to parliamentary rules, others are preserved at the insistence of female MPs who refuse to let these comments be erased from history. Their insistence to keep such remarks in the official record reflects a behavior shift, one where women in parliament are no longer willing to keep quiet and comply when insulted or undervalued.

Linguistic practices also play a crucial role in sustaining or disrupting inequality. The use of masculine generics continues to dominate parliamentary speech, despite clear institutional efforts to promote inclusive alternatives. While the growing presence of gender-inclusive phrases is promising, the fact that female identifiers like "las diputadas" appear mostly within those inclusive constructions, and rarely independently, signals an uneven acknowledgment of women politicians.

Additionally, the institutional framework itself shapes the discursive environment. The regulation that allows offensive speech to be removed from official records was originally designed to maintain

respect and order. However, it now risks hiding those sentences in speeches that reveal persistent gender-based disrespect. This contradiction between keeping order and erasing harmful comments shows a deeper issue in parliamentary culture, where equality may be promoted in theory but not always fully practiced.

Taken together, these findings paint a complex but hopeful picture. The Spanish Parliament has made remarkable progress in closing gender gaps in representation, participation, and visibility. Women now speak more, on more topics, and in increasingly central institutional roles. Yet the path to full equality remains unfinished, and structural progress has not been equally matched by cultural and discursive change. Power still interacts with gender in ways that can diminish voices, question their authority, or marginalize certain forms of expression.

To summarize, the analysis shows that gender equality in political discourse is not something that can be achieved once and then considered finished. While institutional reforms have helped create a more balanced environment, deeper change takes time. It means continuing to question and improve the language, habits, and attitudes that still reflect old power structures. Only then will the *Congreso de los Diputados* become not just a place where women are present, but one where their voices are fully heard and respected.

Chapter 6

Conclusions, Limitations, and Future Work

6.1 Conclusions

At the beginning of this thesis, six hypotheses were proposed to guide the analysis of gendered power dynamics in the Spanish Parliament. After collecting hundreds of documents, segmenting and ordering them, reviewing thousands of interventions, and analyzing data and language patterns, it is time to reflect on what this investigation has uncovered.

The idea that sexism is structurally embedded in parliamentary institutions has data that supports it. In 2003 and 2013, women spoke significantly less than men, not only were they less represented, but those who were present spoke less than their male peers. It was not until 2003 that speech and representation began to align. This imbalance suggests that even when women gain seats, institutional structures continue to limit their ability to participate equally.

Another key takeaway is that sexism does not need to be loud or obvious to be. Sometimes it appears in more subtle forms, such as tone, topic assignment, or language choices. This pattern became clear when analyzing the use of masculine generics, the emotional tone of interventions, and the reactions that female speeches generated. These are not necessarily hostile behaviors, but they reinforce differences in how MP is treated depending on its gender. The idea that sexism often presents itself in an indirect or "benevolent" form is undeniably reflected in the studied data.

Analysis of interruptions brought interesting findings; once normalized by word count, women were more frequently interrupted than men across every year studied. The differences were not always statistically significant at the strictest level, but the trend was consistent. This supports the hypothesis that women are more likely to face resistance, either subtle or explicit, when they speak.

The relationship between language and power was another central piece of this project. Through topic modeling, it was shown that, particularly in the earlier years, women were more likely to speak on topics like education or culture, while money, defense, and foreign affairs were overwhelmingly male-dominated. In 2023, that pattern began to shift. Women participated across all categories and led conversations in themes where they were historically excluded. That change is meaningful, it shows that discursive space can be reclaimed and that gendered topics do not exist.

These patterns reinforce one of the key claims of the thesis, that descriptive representation does not automatically translate to discursive or symbolic equality. Having women in seats does not guarantee their voices are equally welcomed or respected. Representation is a necessary condition, but not a sufficient one.

Beyond the gender analysis, one of the most important contributions of this project is the creation of a structured and centralized relational database of Spanish parliamentary discourse. This database was designed to solve a technical problem, the methodologies used until 2025 to publish parliamentary transcripts make them difficult to work with. The current digital infrastructure of the *Congreso de los Diputados* is not built for deep analysis [17]. The documents are published in separate HTML files, the metadata is incomplete, and the speaker information is scattered, ambiguous, or sometimes simply wrong. No search over all documents contents is enabled, which makes finding concrete information very challenging.

The creation of a database for this project claimed to tackle that problem directly by connecting interventions with speakers' metadata across different terms, cleaning and correcting inconsistencies, and restructuring the information in a normalized format. The result is not a dataset, but an entire infrastructure that enables complex multi-dimensional queries across twenty years of parliamentary discourse. It is now possible to filter directly by gender, group, year, speaker, keyword, and more. And because it is all structured, filtering is fast, reliable, and reproducible.

On a personal level, it is also a reflection of what engineering means. This is a data science project at its core, one that applies engineering principles to structure chaos, build pipelines, and generate meaningful knowledge from raw and messy data. It demonstrates how computational thinking and technical design can be used to ask social questions and to give voice to patterns that might otherwise remain invisible.

In the end, this research shows that change is happening, but it is slow, uneven, and still incomplete. The Spanish Parliament today is different from what it was twenty years ago, women are more present, more vocal, and more central. But presence alone is not enough, until voices are received with equal legitimacy, until language reflects that equality, and until institutional structures support it, the promise of parity remains unfulfilled.

This thesis does not pretend to offer final answers. But it does show that listening closely to language, to participation, to silence, reveals more than just words. It reveals how power works. And through data, code, and care, it also shows how engineering can help us see it more clearly.

6.2 Limitations

This project, like any academic work, has its limitations. Perhaps, the most significant one is the temporal scope, only three years were selected for the analysis. While this allowed for meaningful comparisons across two decades, it also meant leaving out much relevant data. Ideally, the project would include all data available from 2003 to 2023, which would allow for more continuous analysis of change, revealing gradual trends, reactions to political events, or sudden shifts in speech. However, given the size of the original dataset and the computational resources available, this approach was not feasible. Focusing on only three strategically chosen years was a necessary compromise to allow for depth without sacrificing quality.

Another limitation involves the use of sentiment analysis tools. While the methods applied were carefully chosen and validated, sentiment analysis remains a complex task, especially in a political context. Language in parliaments is formal, sometimes ironic, words may sound neutral on the surface while still carrying judgment or vice versa. The models used were good at capturing broad tone patterns, but they sometimes exhibit a large error rate, especially when dealing with emotionally neutral institutional language. For this reason, the sentiment results should be understood as general trends rather than exact emotional readings.

There is also a limitation in terms of the source material, the analysis is based on official transcripts from the *Diario de Sesiones*, which are subject to internal rules that can affect the final content. Specifically, Article 104 of the *Reglamento del Congreso* allows the President to remove offensive remarks from the official record. While this rule was created to preserve institutional respect, it also acts as a form of censorship. Explicitly sexist or disrespectful comments may never appear in the dataset, which means that part of the reality this project aims to study could be missing from the source itself. This kind of institutional bias limits how complete any analysis of language and sexism can be when it depends on official records alone.

6.3 Research transparency

6.3.1 Positional Statement

The researcher is a 22-year-old woman who identifies with left-leaning political values and is a feminist. This personal and political background inevitably shapes how the material is approached and understood. While academic research should aim to be rigorous and as fair as possible, complete neutrality is not something that can be fully achieved.

Knowing this, all efforts were made to build the analysis based on data, not opinions, and to be transparent about methods and interpretations. Sharing this positionality is part of the project's commitment to honesty and awareness, but also a warning for readers, there is bias in this thesis, not intended, but as a human, it is impossible to be fully unbiased.

6.3.2 Use of AI Tools for Writing Assistance

AI-based language tools were used to support the writing of the thesis. These tools were not employed for technical analysis or decision-making, but rather to help improve clarity, coherence, and articulation of ideas. At times, they served as a way to rephrase or better express thoughts that were already present in drafts, especially when trying to communicate complex arguments.

Their use was always guided by the intention to enhance, not replace, human reasoning and interpretation. The content, structure, and conclusions of this work remain the result of personal research, critical thinking, and academic effort. Mentioning the use of AI tools is part of the project's commitment to transparency.

6.4 Future Work

There are many directions this project could take if expanded. The most significant and impactful would be to complete the database with all available legislative years. The potential of this project lies in building a full historical dataset that includes every term since the democratic transition. This would allow for a much more detailed and continuous analysis of parliamentary discourse and its evolution over time.

Such a dataset could then be made publicly available through a research repository and via a search interface. Ideally, users would be able to explore the data using filters such as year, speaker, political party, gender, or topic. They could choose to download specific parts of the dataset, generate summaries, or visualize trends. This would provide journalists, researchers, educators, and all citizens with direct access to institutional documents in a structured and easy format.

The project could go even beyond and integrate AI to enhance the platform. A chatbot interface could help users interact with the data in a conversational way, answering questions like "What was said about healthcare in 2010?" or "Show me all the interventions that mentioned violence against women." It could retrieve quotes, summarize debates, group similar speeches, and even detect patterns like repeated interruptions or the use of offensive language.

Opening up parliamentary data in this way would not only support research but also promote transparency. As of June 2025, public access to this information exists in theory, but in practice, it remains difficult to find and interact with.

Beyond that, the dataset has many possible applications outside gender analysis. It could be used to study polarization, populism, emotional tone in crisis moments, or the evolution of parliamentary language over time. Because the structure was built with flexibility in mind, there is room to adapt it to many different questions and disciplines.

Whether this work continues through future academic projects, public collaborations, or open data initiatives, its potential is clear. The tools are here, what comes next is up to those who choose to build on them.

Bibliography

[1] Barba, A., & Redondo, C. (2024). *Intergenerational evolution of gender bias in Spain: Analysis of values surveys*. Universidad Camilo José Cela. Retrieved May 3rd, 2025, from https://repositorio.ucjc.edu/bitstream/handle/20.500.12020/1606/SI%2013%20-%20Intergenerational%20Evolution%20of%20Gender%20Bias%20in%20Spain_%20Analysis%20of%20Values%20Surveys.pdf?sequence=1&isAllowed=y

[2] Inter-Parliamentary Union. (2016). *Sexism, harassment and violence against women parliamentarians*. Retrieved March 15, 2025, from https://www.ipu.org/resources/publications/issue-briefs/2016-10/sexism-harassment-and-violence-against-women-parliamentarians

[3] Inter-Parliamentary Union. (2018). *Sexism, harassment and violence against women in parliaments in Europe*. Inter-Parliamentary Union. Retrieved October 10th, 2024, from https://www.ipu.org/resources/publications/issue-briefs/2018-10/sexism-harassment-and-violence-against-women-in-parliaments-in-europe

[4] Collier, C. N., & Raney, T. (2018). *Understanding sexism and sexual harassment in politics*. Retrieved February 21, 2025, from https://mlkrook.org/pdf/Collier_Raney_18.pdf

[5] SemEval. (2023). Task 10: *Explainable detection of online sexism*. Retrieved March 4, 2025, from https://arxiv.org/abs/2303.04222

[6] Mishra, S., & Jain, A. (2024). *A deep dive into automated sexism detection using fine-tuned models*. Retrieved March 15, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4791798

[7] Devlin, J., Gupta, P., & Shukla, R. (2021). *Sexism identification in Tweets and Gabs using deep neural networks*. Retrieved April 5, 2025, from https://arxiv.org/abs/2111.03612

[8] Dey, R., & Singh, A. (2024). *Explainable AI discloses gender bias in sexism detection algorithms*. Retrieved April 6, 2025, from https://dl.acm.org/doi/10.1145/3704522.3704524

[9] Krook, M. L. (2017). *Violence against women in politics. Journal of Democracy*, 28(1), 74–88. Retrieved February 13, 2025, from https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1183&context=politicalsciencepub

[10] UN Women. (2024). *Facts and figures: Women's leadership and political participation*. Retrieved February 5, 2025, from https://www.unwomen.org/en/articles/facts-and-figures/facts-and-figures-womens-leadership-and-political-participation

[11] European Institute for Gender Equality. (2023). *Gender Equality Index: Spain*. Retrieved May 16, 2025, from https://eige.europa.eu/gender-equality-index/2023/country/ES

[12] Field, A., & Fakhraei, S. (2021). *Quantifying gender biases towards politicians on Reddit*. Retrieved April 9, 2025, from https://arxiv.org/abs/2112.12014

[13] Jha, R., & Mamidi, R. (2022). *Detecting sexism in YouTube comments using contextualized embeddings*. Retrieved April 9, 2025, from https://arxiv.org/abs/2201.09876

[14] Congreso de los Diputados. (1982). *Reglamento del Congreso de los Diputados*. Retrieved September 26, 2024, from https://www.congreso.es/webpublica/ficherosportal/reglam_congreso.pdf

[15] HYBRIDAS & Dirección General contra la Violencia de Género. (2021). *La violencia política por razón de género: Definición, formas y estrategias para su visibilización.* Retrieved September 26, 2024, from https://violenciagenero.igualdad.gob.es/violenciaPolitica/docs/La_violencia_politica_por_razon_de_genero_HYBRIDAS_DGVG_DEF.pdf

[16] Congreso de los Diputados. (2010). *Recomendaciones para el uso no sexista del lenguaje.* Retrieved November 21, 2024, from https://www.congreso.es/docu/igualdad/Recomendaciones_uso_no_sexista_lenguajeCC.GG..pdf

[17] Congreso de los Diputados. (n.d.). *Repositorio de publicaciones del Congreso y del Senado.* Retrieved March 7, 2025, from https://www.congreso.es/ca/busqueda-de-publicaciones

[18] Congreso de los Diputados. (n.d.). *Listado histórico de diputados.* Retrieved April 4, 2025, from https://www.congreso.es/ca/busqueda-de-diputados

[19] Cerquides Bueno, J., & Mateos Marín, P. (2024). *Spanish Congress parliamentary records (1977–2024)* [Data set]. CSIC. https://doi.org/10.5281/zenodo.11212304

[20] Mateos Marin, P. (2025). *Structured Parliamentary Discourse Dataset from the Spanish Congreso de los Diputados (2003, 2013, 2023)* (Version 1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15495900

[21] Tremlett, G., & Arie, S. (2003, March 29). *Aznar faces 91% opposition to war.* The Guardian. Retrieved from https://www.theguardian.com/world/2003/mar/29/spain.iraq?.com

[22] Wikipedia contributors. (n.d.). *Anti-austerity movement in Spain.* Wikipedia. Retrieved May 27, 2025, from https://en.wikipedia.org/wiki/Anti-austerity_movement_in_Spain

[23] Jones, S. (2023, October 24). *Spain's ruling socialists and leftwing alliance sign coalition deal.* The Guardian. Retrieved from https://www.theguardian.com/world/2023/oct/24/spains-ruling-socialists-and-leftwing-alliance-sign-coalition-deal?.com

[24] Congreso de los Diputados. (2022, November 23). *Diario de Sesiones del Pleno, núm. 224 (14ª legislatura)*. https://www.congreso.es/public_oficiales/L14/CONG/DS/PL/DSCD-14-PL-224.PDF

[25] Congreso de los Diputados. (2024, February 22). *Diario de Sesiones de Comisión, núm. 65 (15ª legislatura)*. https://www.congreso.es/public_oficiales/L15/CONG/DS/CO/DSCD-15-CO-65.PDF

[26] Krook, M. L. (2009). *Quotas for women in politics: Gender and candidate selection reform worldwide*. Oxford University Press. Retrieved from a physical copy, also accessible from: https://doi.org/10.1093/acprof:oso/9780195375671.001.0001

[27] Congreso de los Diputados. (2023). *Reglamento del Congreso de los Diputados* (art. 95). Retrieved June 3rd, 2025, from https://www.congreso.es/cem/t4c7

[28] Congreso de los Diputados. (2012). *Resolución de la Mesa del Congreso de los Diputados, de 26 de junio de 2012, sobre edición electrónica de publicaciones oficiales*. Retrieved June 3rd, 2025, from https://www.congreso.es/cem/20120626-edoficial

[29] Congreso de los Diputados. (2015). *Normas de la Mesa del Congreso de los Diputados, de 20 de enero de 2015, sobre aplicación de la Ley de Transparencia*. Retrieved June 3rd, 2025, from https://www.congreso.es/cem/normtransp